Variationistinen korpustutkimus predikaatin differentiaalisesta lukukongruenssista ja substantiiviluokasta suomen murteissa

Sinnemäki, Kaius

Åbo Akademis förlag
2021

Sinnemäki , K & Haakana , V L J 2021 , Variationistinen korpustutkimus predikaatin differentiaalisesta lukukongruenssista ja substantiiviluokasta suomen murteissa . julkaisussa L M Heikkola , G Paulsen , K Wojciechowicz & J Rosenberg (toim) , Språkets funktion : Juhlakirja Urpo Nikanteen 60-vuotispäivän kunniaksi - Festskrift till Urpo Nikanne på 60-årsdagen - Festschrift for Urpo Nikanne in honor of his 60th birthday . Åbo Akademis förlag , Åbo , Sivut 83-112 .

http://hdl.handle.net/10138/330397

unspecified
publishedVersion

Variationistinen korpustutkimus predikaatin differentiaalisesta lukukongruenssista ja substantiiviluokasta suomen murteissa

Kaius Sinnemäki & Viljami Haakana

Abstract: In this article, we research variation in third person plural verb agreement in Finnish traditional dialects. Unlike in standard Finnish, in traditional dialects the main verb does not always agree in number with a third person plural subject. We approach this variation from a typologically informed variationist perspective using data from The Finnish Dialect Syntax Archive. We model variation in agreement with generalized linear mixed effects modelling and use as predictors speaker's age and sex, length of dependency between the subject and the verb, and the plural subject pronouns *he* and *ne*, part of whose semantic-pragmatic difference lies in animacy. Verb lemma and individual speaker and their dialect are used as random grouping factors. The results suggest that variation in plural agreement is influenced by speaker's sex, dependency length, and the pronoun. The effect of pronoun type on agreement indicates there is some evidence on noun classes in Finnish.[1]

Keywords: number agreement, dependency length, noun class, animacy, corpus research, language typology, Finnish

1 Johdanto

Suomen murteissa ja puhekielessä predikaatit eivät tunnetusti aina kongruoi monikossa monikon kolmannen persoonan subjektin kanssa. Predikaatti esiintyy usein yksikön 3. persoonan muodossa, kuten *lapset/he/ne syö*, eikä monikon 3. persoonan muodossa, kuten *lapset/he/ne syövät* (esim. Helasvuo & Laitinen 2006: 176). Tällaista predikaatin lukukongruenssin vaihtelua on aiemmassa tutkimuksessa käsitelty useista eri näkökulmista. Sitä on tarkasteltu mm. subjektin ja verbin keskinäisen järjestyksen, niiden etäisyyden ja semanttispragmaattisten ominaisuuksien näkökulmista, lausetyypin sekä murteiden ja standardin kirjakielen näkökulmista (mm. G. Karlsson 1966; F. Karlsson 1977; Mielikäinen 1984; Lehikoinen 2000). Eri tekijöiden keskinäistä voimakkuutta ei kuitenkaan ole tutkittu.

Verbin 3. persoonan monikkomuodolla on myös muita kuin monikkokongruenssiin liittyviä funktioita. Murteissa sitä käytetään mm. luomaan tekstuaalista koheesiota (Paunonen 1995) ja lisäksi sillä on impersonaalisuuteen liittyviä funktioita (esim. Forsberg ym. 2009). Joissain murteissa subjektiton monikkokongruenssi (*ø lähtivät*) voi olla lähes täydennysjakaumassa pronominisubjektin kanssa ilman monikkokongruenssia (*he lähti*) (G. Karlsson 1966; Forsberg

[1] Kiitämme Kielipankkia aineiston jakamisesta ja Hanna Westerlundin merkittävää osuutta asian järjestelyissä. Lisäksi kiitämme useita kollegoita tuesta, keskusteluista ja vinkeistä tutkimusprosessin varrella, nimeltä mainiten: Francesca Di Garbo, Ulla-Maija Forsberg, Nobufumi Inaba, Fred Karlsson, Eri Kashima, Leena Manninen, Ricardo Napoleão de Souza, Mikael Parkvall, Mari Siiroinen ja Maria Vilkuna. Lisäksi kiitämme artikkelin kahta arvioijaa. Kiitämme myös yleisöä tutkimustamme koskevista hyödyllisistä kommenteista konferenssissa 95th Annual Meeting of the Linguistic Society of America, jossa esittelimme tutkimustamme. Tämä tutkimus on saanut rahoitusta Euroopan tutkimusneuvostolta (ERC) Euroopan unionin Horisontti 2020 -tutkimus- ja innovaatio-ohjelmasta (sopimus nro 805371).

ym. 2009). Näistä verbin 3. persoonan monikkomuotoon ja sen funktioon ja jakaumaan liittyvästä variaatiosta huolimatta monikkomuotoa käytetään kuitenkin myös kongruenssifunktiossa, mikä on tämän tutkimuksemme keskiössä.

Lähdemme siitä variationistisesta oletuksesta, että predikaatin lukukongruenssin kaltaiseen variaatioon voi vaikuttaa usea tekijä. Variaatio voi liittyä kielen rakenteellisiin tekijöihin, kuten subjektin etäisyyteen verbistä, tai sosiolingvistisiin tekijöihin, kuten puhujan murteeseen, tai mahdollisesti sekä rakenteellisiin että sosiolingvistisiin tekijöihin. Tavoitteemme on selvittää laskennallisesti korpustutkimuksen näkökulmasta, mitkä rakenteelliset ja sosiolingvistiset tekijät vaikuttavat predikaatin lukukongruenssin variaatioon suomen murteissa, kun eri tekijät otetaan huomioon samassa tilastollisessa mallissa. Kiinnitämme erityisesti huomiota siihen, vaihteleeko lukukongruenssi nominisubjektin tyypin mukaan, koska aihe on typologisen tutkimuksen näkökulmasta ajankohtainen ja kiinnostava ja lisäksi se on jäänyt suomalaisessa tutkimuksessa vähälle huomiolle (mutta ks. Mielikäinen 1984).

Selvyyden vuoksi teemme aluksi kolme terminologista erottelua. Puhumme (luvun) inkongruenssista silloin, kun predikaatti esiintyy yksikkömuodossa monikollisen 3. persoonan subjektin kanssa. Nämä ovat tapauksia *lapset/he/ne syö*. Puhumme (luvun) ilmikongruenssista tai monikkokongruenssista silloin, kun predikaatti esiintyy monikkomuodossa monikollisen 3. persoonan subjektin kanssa. Nämä ovat tapauksia *lapset/he/ne syövät*. Kolmanneksi nimitämme tätä vaihtelua yleisesti ottaen differentiaaliseksi lukukongruenssiksi. Kieliopillisten kategorioiden differentiaalinen merkintä on tyypillistä maailman kielissä. Esimerkiksi objektin sijanmerkintä on usein differentiaalista siten, että vain osa verbin objekteista on ilmimerkitty sijalla ja osa on nollamerkitty, usein liittyen tarkoitteen semanttisiin tekijöihin, kuten elollisuuteen ja määräisyyteen (de Swart 2007; Sinnemäki 2014).

Aiemman tutkimuksen perusteella lukukongruenssin vaihteluun vaikuttavat ainakin seuraavat sosiolingvistiset tekijät. Yleiskielessä ilmikongruenssi on normi jo Agricolalla eli vanhan kirjasuomen ajoilta (Häkkinen 1994: 344–347). Inkongruenssin juuret ovat puolestaan murteissa, joista ne ovat alkaneet levitä myös yleiskieleen (mm. Mielikäinen 1984; Lehikoinen 2000). Murteiden välillä on myös eroja lukukongruenssin esiintymisessä. Inkongruenssi on yleisintä hämäläis- ja savolaismurteissa sekä etelä-, keski- ja pohjoispohjalaismurteissa, kun taas ilmikongruenssi on yleisintä lounaismurteissa, peräpohjalaisissa murteissa ja kaakkoismurteissa (esim. G. Karlsson 1966). Aivan selvää systematiikkaa lukukongruenssin maantieteellisissä jakaumissa ei silti näytä esiintyvän. Sekä ikä että sukupuoli näyttävät vaikuttavan kongruenssin vaihteluun myös. 1970-luvun kaupunkilaisilta kerätyn aineiston perusteella inkongruenssi on selvästi yleisempää nuorten puheessa verrattuna iäkkäämpien puheeseen, kun taas ilmikongruenssi on yleisempää naisilla verrattuna miehiin, etenkin ylemmissä sosiaaliluokissa (Mielikäinen 1984: 163–164).

Lukukongruenssin vaihteluun vaikuttaa aiemman tutkimuksen mukaan myös seuraavat kielen rakenteelliset tekijät. Inkongruenssi on yleistä tietyillä verbeillä, etenkin *olla*-verbillä. Ilmikongruenssi näyttää esiintyvän etenkin silloin kun subjektin ja verbin etäisyys on suuri tai kun lauseessa ei esiinny ilmisubjektia lainkaan (so. *lapset leikkii* tai *ø leikkivät* mutta ei

niinkään *lapset leikkivät*; esim. Mielikäinen 1984; Forsberg ym. 2009). Tosin inkongruenssia esiintyy ainakin 1990-luvun lopun ylioppilasaineissa sitä enemmän, mitä etäämmällä subjekti sijaitsee verbistä (Lehikoinen 2000: 257). Ilmikongruenssia esiintyy todennäköisemmin, kun subjekti edeltää verbiä (so. *lapset leikkivät*) ja inkongruenssi puolestaan todennäköisemmin, kun subjekti seuraa verbiä (so. *leikkii lapset*) (F. Karlsson 1977: 371–372; Lehikoinen 2000: 247). Myöntölauseissa ilmikongruenssi esiintyy lisäksi tavallisemmin kuin kieltolauseissa (so. *lapset leikkivät* vs. *lapset ei leiki*; G. Karlsson 1966; Lehikoinen 2000: 255–258). Subjektin semanttiset ominaisuudet näyttävät myös vaikuttavan kongruenssin vaihteluun siten, että ilmikongruenssi esiintyy etenkin henkilösubjektien kanssa, joissain aineistoissa jopa lähes pakollisesti: so. *lapset leikkivät* vs. *puut huojuu*. (Mielikäinen 1984: 166–172).

Suomenkielisessä tutkimuksessa ei tietääksemme aiemmin ole kuitenkaan huomioitu sitä typologisesta tutkimuksesta kumpuavaa seikkaa, että jos kongruenssi vaihtelee nominin ominaisuuksien mukaan, tällöin puhutaan määritelmällisesti substantiiviluokasta tai kieliopillisesta suvusta (esim. Di Garbo, Olsson & Wälchli 2019). Differentiaalinen lukukongruenssi ei itsessään vielä ole merkki substantiiviluokan esiintymisestä, mutta jos predikaatin lukukongruenssi vaihtelee tilastollisesti nominisubjektin mukaan, tätä voidaan pitää evidenssinä substantiiviluokan kehittymisestä kieleen. Usein tällainen vaihtelu liittyy elollisuuteen ja voidaan sanoa, että kongruenssin todennäköisyys kasvaa, mitä ylempänä nominireferentti on ns. elollisuushierarkiassa (1) (esim. Corbett 2010: 91–93).

(1) Elollisuushierarkia (Corbett 2010: 92)
1 > 2 > 3 > sukulainen > muut henkilöt > muut elolliset > eloton

Esimerkiksi papualaisessa munan kielessä verbi esiintyy monikkomuodossa ihmistarkoitteisten monikkosubjektien kanssa (2a), yksikkömuodossa elottomien monikkosubjektien kanssa (2b) ja varioi muiden kuin ihmisiin viittaavien elollisten monikkosubjektien kanssa (2c). Samantyylistä vaihtelua esiintyy myös mm. turkissa ja persiassa (Bamyacı, Häussler & Kabak 2014; Bayanati & Toivonen 2019).

(2) a. *ihintu-umu* *o-kala-amu*
2-PL 2-mennä-PL
‘te menette’ (van den Berg 1989: 51)
b. *bara-hi-no* *no-hali*
tavara-PL-hänen 3SG.REAL-kallis
‘hänen tavaransa ovat kalliita’ (van den Berg 1989: 52)
c. *o* *kadadi-hi* *no-rato-mo* / *do-rato-mo*
ART eläin-PL 3SG.REAL-saapua-PRF / 3PL.REAL-saapua-PRF
‘eläimet ovat saapuneet’ (van den Berg 1989: 52)

Kongruenssin vaihtelun ei kuitenkaan tarvitse liittyä nominin semantiikkaan ollakseen merkki substantiiviluokasta: riittää, että vaihtelu on systemaattista. Substantiiviluokat mm.

bantukielissä eivät määräydy kovin selvästi semanttisin perustein. Mutta jos kongruenssin vaihtelu liittyy nominin semantiikkaan, pohjimmiltaan erottelut liittyvät yleensä elollisuuteen (Dahl 2000). Tämän vuoksi on tärkeää kiinnittää huomiota subjektinominin tyyppiin.

Mutta jos kerran monikon kolmannen persoonan pronominin valinta riippuu elollisuudesta, eikö pronomineja voitaisi jo itsessään pitää merkkinä substantiiviluokasta, kuten artikkelimme yksi arvioija esittää? Näin tosiaan voisi ajatella. Monet kongruenssitutkijat ovat sitä mieltä, että jos yksi anaforinen pronomini esiintyy esimerkiksi elollisten korrelaattien kanssa (esim. *lapsi … hän*) ja toinen elottomien korrelaattien kanssa (esim. *auto … se*), tämä on jo esimerkki kongruenssista ja substantiiviluokan (tai kieliopillisen suvun) esiintymisestä kielessä (Corbettia 2013 mukaillen). Tässä mielessä esimerkiksi englannissa esiintyy kieliopillinen suku, josta keskeisenä evidenssinä ovat yksikön kolmannen persoonan pronominit *he* ja *she*. Vastaavasti suomen pronominit *hän/he* vs. *se/ne* voitaisiin analogisesti analysoida evidenssinä substantiiviluokan esiintymisestä ainakin yleiskielessä. Siinä missä englannin sukuun viittaavat pronominit analysoidaan käytännössä aina evidenssinä kieliopillisesta suvusta, tutkimuskirjallisuudessa suomen kolmannen persoonan pronomineja ei käytännössä koskaan analysoida evidenssinä substantiiviluokasta. Tämä käytäntö on menetelmällisesti ristiriitainen ja vaatiikin korjausta. Tässä artikkelissa emme kuitenkaan pyri analysoimaan kolmannen persoonan pronominien semanttispragmaattisia ominaisuuksia, koska aiheesta on jo olemassa paljon ansiokasta tutkimusta fennistiikassa (mm. Laitinen 2005; Priiki 2017). Sen sijaan keskitymme tarkastelemaan differentiaalista lukukongruenssia, sillä jos se liittyy nominisubjektin tyyppiin suomen murteissa, tämä on pelkkien pronominien sijaan paljon vahvempaa evidenssiä substantiiviluokan esiintymisestä.

Käytämme aineistona Kielipankissa vapaasti saatavilla olevaa Lauseopin arkiston murrekorpusta (Turun yliopisto, kieli- ja käännöstieteiden laitos & Kotimaisten kielten keskus 1985). Aineistoon on annotoitu tietoa puhujan iästä, syntymäajasta, sukupuolesta ja murteesta sekä runsaasti kieliopillista tietoa, mm. kunkin sanan sanaluokka, lemma ja taivutuskategoriat.

Mallinnamme kongruenssin esiintymistä tilastollisesti yleistetyllä lineaarisella sekamallilla, jota käytetään usein variationistisessa tutkimuksessa.[2] Näiden mallien avulla on aiemmin tarkasteltu mm. yksikön kolmannen persoonan *s*-kongruenssia englannissa (Gries & Hilpert 2010). Selittävinä tekijöinä mallinnamme seuraavia neljää piirrettä: 1. subjektin ja predikaatin välinen etäisyys, jota nimitämme nykytutkimuksen mukaisesti dependenssin pituudeksi, 2. subjektipronominin valinta, 3. puhujan ikä ja 4. puhujan sukupuoli. Subjektipronominin *he* vs. *ne* semanttispragmaattinen ero liittyy ainakin osittain elollisuuteen, tarkemmin ottaen ihmissubjekteihin. Näillä pronomineilla on muitakin funktioita, *ne*-pronominilla mm. impersonaalisuuteen liittyviä ja *he*-pronominilla logoforisuuteen liittyviä funktioita (mm. Mielikäinen 1984; Forsberg ym. 2009; Priiki 2017). Riippumatta näiden pronominien tarkasta semanttisesta erosta niiden käyttö on saattanut alkaa eriytyä kongruenssin suhteen, kuten esimerkiksi Priiki (2017: 80) uumoilee. Verbin lemmaa ja murteita mallinnamme ns.

[2] Ellonen (2006) tarjoaa lyhyen suomenkielisen esityksen näihin malleihin yhteiskuntatieteissä. Tervola (2019) on hiljattain soveltanut niitä suomenkielisen aineiston tutkimukseen.

satunnaistermeinä, niin että jokaiselle verbin lemmalle ja jokaiselle murteelle estimoidaan oma satunnaisvakionsa (engl. *random intercept*), joka kuvaa kunkin verbin ja murteen vaikutusta kongruenssiin (murteista ja niiden luokituksesta ks. luku 2.1; mallintamisesta ks. luku 2.2).

Sosiolingvistisen tutkimuksen perusteella naisten puhe on tyypillisesti yleiskielisempää kuin miesten. Lisäksi puhujien ikä vaikuttaa usein kielelliseen variaatioon. Koska monikon kolmannen persoonan lukukongruenssi on kirjakielen piirre, voisi olettaa, että naiset käyttäisivät sitä miehiä useammin myös murteissa. Lauseopin arkiston murrekorpus on kuitenkin tarkoituksella kerätty niin, että puhujat edustaisivat melko homogeenistä joukkoa, jotta haastateltavien väliset erot edustaisivat vain murre-eroja (mm. Ikola 1985; ks. luku 2.1). Aineiston koostumuksen perusteella on siis epäselvää, missä määrin mm. yleiskielisyydet olisivat vaikuttaneet haastateltavien kielenkäyttöön, koska ilmeisiä yleiskielisyyksiä ei olisi huolittu aineistoon. Vaikka murreaineistosta on pyritty poistamaan ilmeisen yleiskielinen aines, on silti perusteltua selvittää haastateltavan sukupuoleen ja myös ikään mahdollisesti liittyviä eroja monikkokongruenssin käytössä. Esimerkiksi sukupuolierot voivat olla varsin hienoisia ja siten vaikeasti havaittavissa ilman määrällistä analyysia. Haastateltavan sukupuoli ja ikä on myös annotoitu aineistoon, joten niiden vaikutusta on helppo tarkastella tilastollisen mallin avulla.

Tulosten perusteella kongruenssin vaihteluun vaikuttavat merkittävällä tavalla mallinnetut kielen rakenteelliset tekijät, eli verbin lemma, dependenssin pituus ja subjektipronomini (*he* vs. *ne*). Lisäksi vaihtelua selittää yksittäisestä puhujasta ja hänen murteestaan johtuva vaihtelu sekä vähäisessä määrin puhujan sukupuoli mutta ei hänen ikänsä. Sukupuolen vaikutus voi viitata haastateltavien naisten yleiskielisempään puheeseen verrattuna miehiin. Dependenssin pituuden vaikutus kongruenssin esiintymiseen selittyy pääosin ekonomialla. Pronominin vaikutus kongruenssin vaihteluun näyttää antavan tukea substantiiviluokan esiintymiselle suomen murteissa. Myös sen kehittymiseen on nähtävästi voinut vaikuttaa ekonomia.

Artikkelin rakenne on seuraava: luvussa 2 esittelemme käyttämämme aineiston ja menetelmät, luvussa 3 esitämme korpusanalyysin tulokset, luvussa 4 keskustelemme artikkelin tuloksista ja jatkotutkimuksen keskeisistä kysymyksistä ja luvussa 5 esitämme lyhyet johtopäätökset.

2 Aineisto ja menetelmät

2.1 Aineisto ja sen analyysi

Aiemman tutkimuksen perusteella verbin 3. persoonan lukukongruenssin vaihtelua esiintyy etenkin murteissa. Tässä tutkimuksessa tarkastelemme lukukongruenssin vaihtelua Lauseopin arkiston murrekorpuksen aineistossa (Turun yliopisto, kieli- ja käännöstieteiden laitos & Kotimaisten kielten keskus 1985). Se sopii tarkoitukseemme hyvin, koska se sisältää aineistoa perinteisistä murteista ja siihen on annotoitu sekä kieliopillista tietoa että tietoa yksittäisistä puhujista. Lauseopin arkiston aineisto on kerätty 1950–1970-luvuilla, ja se sisältää nauhoitettua puhetta yli 100 haastateltavalta, kattaen pääasiassa kerrontaa henkilöhistoriasta, tavoista ja työstä. Haastateltavat ovat syntyneet pääosin 1880- ja 1890-luvuilla ja ovat näin

ollen suurin piirtein samanikäisiä. Taustaltaan he ovat kouluttamattomia maaseudun asukkaita, joiden puheeseen kirjakielen normit eivät ole vaikuttaneet (ks. Ikola 1985). Näiden kriteerien avulla on pyritty saamaan edustava joukko haastateltavia, joiden väliset erot heijastaisivat aidosti vain murre-eroja. Kuvaamme seuraavassa lyhyesti aineiston analyysia.

Etsimme murrekorpuksesta automaattisesti ensin sellaiset tapaukset, joissa monikollinen pronomini *he* tai *ne* esiintyy nominisubjektina nominatiivissa ja niitä seuraa predikaattiverbi. Partitiivisubjektit jätimme näin ollen analyysin ulkopuolelle. Predikaattiverbiksi analysoimme kaikki finiittiverbit, mukaan lukien liittopredikaatit. Liittopredikaateista kongruenssianalyysiin on otettu vain sen finiittinen osa. Keskitymme ainoastaan myöntölauseisiin, koska kieltoverbin *ei* lukukongruenssi on murteissa tunnetusti harvinaista (G. Karlsson 1966: 21; Mielikäinen 1984: 166). Koska aineistoon ei ole annotoitu pääsanan ja dependenttien välisiä dependenssisuhteita, rajasimme subjektin ja predikaattiverbin etäisyyden tarkastelun tapauksiin, joissa niiden etäisyys on 0–4 sanaa (eli jolloin niiden välissä esiintyy 0–4 sanaa pois lukien tauot).[3] Olisi kyllä mahdollista analysoida tapaukset, joissa subjektin ja verbin etäisyys olisi suurempi kuin neljä sanaa. Rajasimme tarkastelun kuitenkin pienempiin etäisyyksiin, jotta työmäärä pysyisi kohtuullisena, koska subjektin ja predikaattiverbin dependenssisuhde oli analysoitava osittain manuaalisesti. Tulosten perusteella dependenssin pituuden vaikutus kongruenssin vaihteluun voidaan joka tapauksessa osoittaa jo pienilläkin dependenssipituuksilla. Rajasimme ulkopuolelle myös tapaukset, joissa predikaattiverbi edeltää subjektipronominia, koska sanajärjestyksen ottaminen huomioon olisi entisestään lisännyt manuaalisen työn määrää (ks. alla). Lisäksi typologisen evidenssin perusteella kielissä, joissa subjektin ja predikaatin sanajärjestys vaikuttaa monikkokongruenssiin, ilmikongruenssi on todennäköisempää silloin, kun subjekti esiintyy lauseessa ennen predikaattia (Greenberg 1966: 74), joten postverbaaliset subjektit on perusteltua jättää tutkimuksen ulkopuolelle.

Seuraavaksi analysoimme manuaalisesti subjektin ja predikaattiverbin dependenssisuhteen selvittääksemme niiden välisen dependenssin pituuden. Poistimme tarkastelusta sellaiset tapaukset, joissa subjektipronominin ja predikaatin välissä oli mikä tahansa sana, jonka syntaktinen funktio oli nominisubjekti sillä oletuksella, että verbi ei kongruoisi välttämättä poimimamme pronominin kanssa vaan sitä lähempänä esiintyvän toisen nominisubjektin kanssa. Esimerkiksi Ikaalisten aineistossa lauseen (3) analysoimme niin, että predikaattiverbin *hommah* lähimpänä subjektina on *noi kauppioot*, vaikka pronomini *ne* on myös annotoitu nominisubjektiksi. Periaatteidemme mukaisesti tämä lause rajattiin tarkastelun ulkopuolelle.

(3) *ne tooh noi kauppioot hommah niinku noi ölijykkit tänne.*

[3] Määritimme dependenssin pituudeksi subjektin ja predikaattiverbin välissä olevien sellaisten sanojen lukumäärän (+1), jotka eivät koostuneet pelkistä vinoviivoista eli eivät olleet taukoja. Vinoviivat mukaan luettuna välissä sai olla korkeintaan neljä sanaa, mutta ne jätettiin huomiotta dependenssin pituutta laskettaessa. Aineistossa ei siis ole mukana esimerkiksi tapauksia, joissa pronominin ja predikaatin välissä olisi neljä puhuttua sanaa ja niiden lisäksi vielä tauko.

Poistimme analyysista myös tapaukset, joissa verbi esiintyi jossain muussa taivutusmuodossa kuin yksikön tai monikon kolmannen persoonan muodossa. Esimerkiksi Ilomantsin murrenäytteessä monikon kolmannen persoonan subjektin kanssa käytetään myös verbistä passiivimuotoa, kuten lauseessa (4) (ks. myös Mielikäinen 1984).

(4) *ne korpselällä käytih aena*

Pyrimme analyysissa luottamaan mahdollisimman pitkälle Lauseopin arkiston annotointiin ja tekemään analyysimme automaattisesti kieliteknologisin menetelmin. Manuaalisessa analyysissa otimme kuitenkin joissakin tapauksissa kantaa aineiston annotointiin ja korjasimme selvät virheet. Esimerkiksi Pohjaslahden aineistossa lauseessa (5) predikaattiverbi *veisaa* oli analysoitu yksikön 1. persoonan imperfektimuodoksi.

(5) *mutta palijo ne jo veisaa / uusia / siionivvirsiä*

Korjasimme tämän verbin muotoon yksikön kolmannen persoonan preesens ja otimme sen mukaan lopulliseen aineistoon. Joissain yksittäisissä tapauksissa verbin perusmuoto on aineistossa epäselvä, mutta verbin kieliopilliset kategoriat on silti analysoitu morfologisesti. Esimerkiksi Punkalaitumen aineistossa lauseessa (6) verbin *viottee* perusmuoto on epäselvä. Kelpuutimme tällaiset tapaukset mukaan, mutta niiden lemmaksi merkitsimme kullekin oman tunnuksen a1, a2, … a10 (tapauksia oli yhteensä 10).[4]

(6) *ne / viottee*

Puhujia koskevat sosiolingvistiset tiedot, kuten haastateltavan murteen, otimme Lauseopin arkistosta sellaisenaan. Lisäksi käytimme syntymäaikaa, haastattelun päivämäärää, sukupuolta ja paikkakuntaa koskevia tietoja hyväksi erotellaksemme toisistaan yksittäiset haastateltavat. Aineistossa yksittäistä rymättyläläistä murteenpuhujaa on haastateltu kahteen otteeseen eri vuosina, mutta yhdistimme tämän yhden puhujan tuotokset samaan yksilötunnukseen.

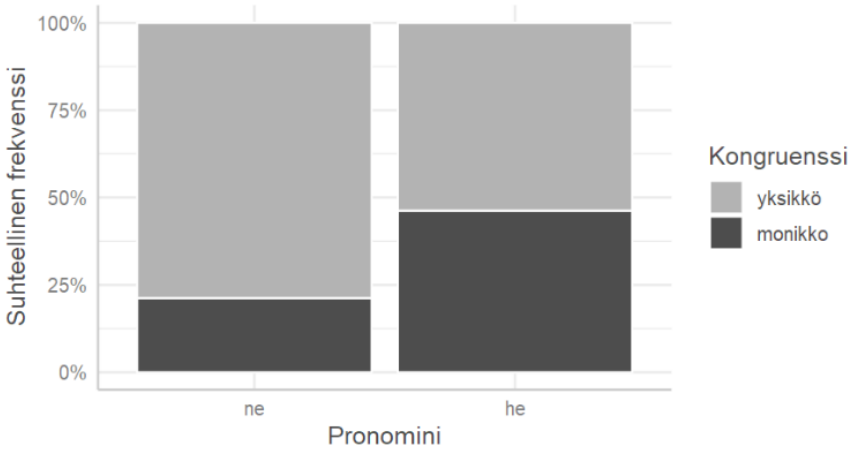
Murretutkimuksessa suomen murteet on jaoteltu erilaisiin pääryhmiin ja niiden alaryhmiin. Standardinomaisin jaottelu luokittelee murteet kahdeksaan pääryhmään (mm. Itkonen 1983, 1989).[5] Lauseopin arkiston murrekorpuksessa murteet on jaettu kuuteen pääryhmään eli murrealueeseen ja hienovaraisemmin 25 alaryhmään eli murreryhmään. Murrealueiden ja murreryhmien jako on esitetty taulukossa 1, josta käy ilmi myös niistä käytetyt lyhenteet.

[4] Analysoimamme aineisto sekä käyttämämme Python- ja R-algoritmit ovat saatavilla artikkelin elektronisessa oheismateriaalissa osoitteessa: https://version.helsinki.fi/gramadapt/verbinlukukongruenssi2021.
[5] Tämä luokittelu ja sitä havainnollistava kartta ovat saatavilla Kotimaisten Kielten Keskuksen verkkopalvelusta: https://www.kotus.fi/kielitieto/murteet/suomen_murteet.

Taulukko 1. Lauseopin arkiston murrekorpuksen murrealueet ja murreryhmät lyhenteineen

Murrealue		Murreryhmä	
Nimi	Lyhenne	Nimi	Lyhenne
Pohjalaismurteet	POH	Etelä-Pohjamaa	PohE
		Keski-Pohjamaa	PohK
		Pohjois-Pohjamaa	PohP
		Länsipohja	LänP
		Peräpohjola	PerP
Savolaismurteet	SAV	Eteläinen Keski-Suomi	KesE
		Läntinen Keski-Suomi	KesL
		Pohjoinen Keski-Suomi	KesP
		Etelä-Savo	SavE
		Pohjois-Savo	SavP
		Pohjois-Karjala	KarP
		Kainuu	Kai
Hämäläismurteet	HÄM	Etelä-Häme	HämE
		Kaakkois-Häme	HämK
		Pohjois-Häme	HämP
		Pohjois-Satakunta	SatP
		Kymenlaakso	Kym
Lounaiset välimurteet	LVÄ	Etelä-Satakunta	SatE
		Länsi-Satakunta	SatL
		Varsinais-Suomen ylämaa	VarY
		Länsi-Uusimaa	VarU
Lounaismurteet	LOU	Eteläinen Varsinais-Suomi	VarE
		Pohjoinen Varsinais-Suomi	VarP
Kaakkoismurteet	KAA	Etelä-Karjala	KarE
		Keski-Karjala	KarK

Lopullinen analysoitu aineisto koostuu 8961 lauseesta. Aineistossa esiintyy inkongruenssia yhteensä 6959 (78 %) tapauksessa ja ilmikongruenssia 2002 (22 %) tapauksessa. Verbin lemmalla on vinoutunut jakauma: 43 %:ssa tapauksista esiintyy *olla*-verbi. Koska analyysimme sisälsi myös liittopredikaatit, kaikki nämä tapaukset eivät kuitenkaan edusta kopulaa. Subjektipronomineista *ne*-pronomineja on 8543 (95 %) ja *he*-pronomineja 418 (5 %). Puhunnoksista lähes tarkalleen puolet on tuottanut nainen ja puolet mies. Haastateltavat olivat haastatteluhetkellä keskimäärin 80-vuotiaita (vaihteluväli 43–94 vuotta).



Kuvio 1. Kongruenssin suhteellinen frekvenssi monikon 3. persoonan subjektipronomineilla

Differentiaalinen lukukongruenssi näyttää alustavasti riippuvan ainakin subjektipronominin valinnasta. Kun asiaa testataan Fisherin tarkan testin avulla, ilmikongruenssi on noin 3,2 kertaa todennäköisempi *he*-pronominin yhteydessä verrattuna *ne*-pronominiin ($p < 0{,}001$; ks. kuvio 1). Jotta lukukongruenssiin vaikuttaviin eri tekijöihin saataisiin parempi selvyys, on sen jakaumaa mallinnettava monimuuttujamallilla.

2.2 Tilastollinen mallintaminen

Selvitämme lukukongruenssin vaihtelua mallintamalla siihen vaikuttavia tekijöitä yleistetyllä lineaarisella sekamallilla (engl. *generalized linear mixed effects model*). Tällaisia malleja käytetään yleisesti variationistisessa korpustutkimuksessa sekä kokeellisessa tutkimuksessa, koska niiden avulla mallissa voidaan ottaa huomioon mm. puhujien keskinäinen vaihtelu tai vaikkapa yksittäisestä lekseemistä johtuva vaihtelu käsittelemällä niitä niin sanottuina satunnaistermeinä (ks. Baayen, Davidson & Bates 2008; Gries & Hilpert 2010; Baayen 2013). Seuraavassa kuvaamme lyhyesti tilastollista malliamme.

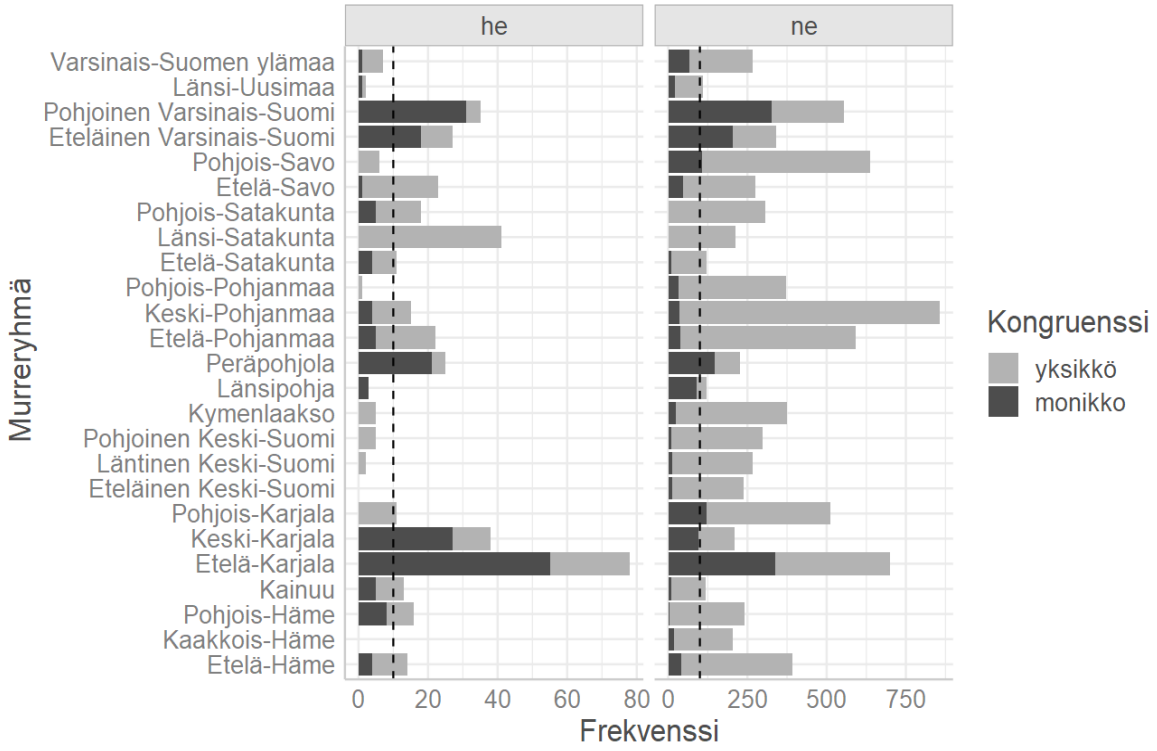
Mallimme selitettävänä muuttujana on kolmannen persoonan kongruenssi predikaattiverbissä. Tämä muuttuja saa kaksi mahdollista arvoa: ”sg” ja ”pl”. Näistä ensimmäinen tarkoittaa, että predikaattiverbissä esiintyy yksikkömuoto monikkosubjektin kanssa (*ne leikkii*) ja jälkimmäinen sitä, että predikaattiverbissä esiintyy monikkokongruenssi monikkosubjektin kanssa (*ne leikkivät*). Koska selitettävä muuttuja on binaarinen, mallinnamme sen jakaumia logistisella regressiolla käyttäen yleistettyä lineaarista sekamallia.

Selittävinä tekijöinä käytämme neljää muuttujaa. Kielen rakenteellisina selittävinä tekijöinä käytämme (i) subjektin ja predikaattiverbin välistä etäisyyttä eli dependenssin pituutta ja (ii) subjektinominin tyyppiä. Dependenssin pituus saa arvoja välillä 0 ja 4. Arvo 0 esiintyy, kun subjekti esiintyy välittömästi ennen verbiä. Etäisyys kasvaa sitä mukaa, mitä useampia sanoja esiintyy subjektin ja verbin välissä. Subjektinomini koodataan tässä tutkimuksessa siten, että se saa arvoksi ”he” tai ”ne” sen mukaan, kumpi pronomini esiintyy verbin subjektina. Lisäksi mukana on kaksi sosiolingvististä selittävää tekijää: (i) puhujan sukupuoli ja (ii) puhujan syntymävuosi. Puhujan sukupuoli on binaarinen muuttuja, jolla on arvoina ”nainen” ja ”mies”.

Lauseopin arkistoon on merkitty puhujan ikä sekä syntymäaika päivämääränä. Koska haastattelut on tehty 18 vuoden aikana (vuosien 1957–1974 välillä), arvelemme, että puhujan iän sijaan hänen syntymäaikansa voisi olla parempi indikaattori ajan vaikutuksesta kongruenssin vaihteluun. Otimme kunkin puhujan syntymäajasta hänen syntymävuotensa ja aloitimme ajanlaskun varhaisimmasta vuodesta. Näin syntymävuosi 1865 koodattiin arvolla 1 ja myöhäisin syntymävuosi 1916 koodattiin arvolla 52.

Selittävien tekijöiden lisäksi käytimme mallissa neljää satunnaistermiä: (i) haastateltavaa, (ii) murreryhmää, (iii) murrealuetta ja (iv) verbin perusmuotoa. Yksittäisen haastateltavan puhunnokset eivät ole riippumattomia toisistaan, joten huomioimme tämän rakentamalla yksittäisen haastateltavan malliin satunnaisterminä. Koska kunkin puhujan näyte edustaa tiettyä murreryhmää, huomioimme tämän hierarkian kiinnittämällä kunkin haastateltavan näyte tiettyyn murreryhmään ja murreryhmän murrealueeseen (ks. 7). Verbin perusmuoto eli lemma on tärkeä ottaa huomioon, koska verbien frekvenssi vaihtelee aineistossa erittäin paljon. Aineistossa esiintyy yhteensä 785 verbiä, mutta niiden esiintymistä 43 %:ssa esiintyy *olla*-verbi ja 52 %:a verbin lemmoista ovat ns. hapax legomena-tapauksia, eli ne esiintyvät aineistossa vain kerran. Koska yksittäisen verbin esiintymä ei välttämättä ole riippumaton verbin muista esiintymistä vaan voi riippua verbin frekvenssistä, huomioimme tämän rakentamalla verbin lemman malliin satunnaisterminä.

Aineiston mukaan sekä lukukongruenssi että subjektipronomini vaihtelevat aineistossa murreryhmittäin (kuvio 2). Monikkokongruenssia esiintyy suhteellisen paljon etenkin Eteläisen Varsinais-Suomen (VarE) ja Pohjoisen Varsinais-Suomen (VarP) murreryhmissä (so. lounaismurteissa), Etelä-Karjalan (KarE) ja Keski-Karjalan (KarK) murreryhmissä (so. kaakkoismurteissa) sekä Länsi-Pohjan (LänP) ja Peräpohjolan (PerP) murreryhmissä. Kaikissa aineiston murreryhmissä *ne*-pronominin esiintymiä on yli sata (pystysuora katkoviiva), mutta muutamissa murreryhmissä *he*-pronominin esiintymiä on alle kymmenen (pystysuora katkoviiva) ja kahdessa murreryhmässä niitä ei esiinny lainkaan (KesE = Eteläinen Keski-Suomi ja HämK = Kaakkois-Häme). Vaikuttaa myös siltä, että alueet, joilla *he*-pronominia esiintyy paljon, myös monikkokongruenssia esiintyy paljon, joskaan aivan selvää tämä ei ole.



Kuvio 2. Lukukongruenssin absoluuttiset frekvenssit eri pronomineilla murreryhmittäin

Tämän vaihtelun vuoksi mallinnamme lukukongruenssia kahdella tavalla. Ensin mallinnamme lukukongruenssin vaihtelua koko aineistossa käyttäen rakenteeltaan yksinkertaisinta monitasomallia. Tässä ns. satunnaisvakiomallissa estimoimme verbin lemmalle ja yksittäiselle haastateltavalle omat satunnaisvakionsa (engl. *random intercept*). Nämä termit kuvaavat verbin lemman ja haastateltavan ja hänen murteensa vaikutusta kongruenssin vaihteluun. Toisessa mallissa huomioimme, että pronominin vaikutus lukukongruenssiin voi vaihdella eri murreryhmissä. Tämä ns. satunnaiskerroinmalli on muutoin samanlainen satunnaisvakiomallin kanssa, mutta rakennamme siihen satunnaiskertoimen, joka ottaa huomioon pronominin vaikutuksen lukukongruenssiin murteittain ja riippuen myös haastateltavasta.[6] Tässä mallinnuksessa rajaamme tarkastelun murreryhmiin, joissa *he*-pronomineja esiintyy vähintään kymmenen (ks. kuvio 2). Tällä tavalla pyrimme saamaan aineistosta ensin kokonaiskuvan satunnaisvakiomallin avulla ja sitten rajaamaan tarkastelun niihin tapauksiin, joissa aineistoa on riittävästi myös kompleksisemmalle satunnaiskerroinmallille.

Tilastolliseen mallintamiseen käytimme avoimesti saatavilla olevaa ohjelmointiympäristöä R (R Core Team 2018) ja sen pakettia `lme4` (Bates, Maechler, Bolker & Walker 2015).[7] Satunnaisvakiomallille käytimme `lme4`:n notaatiota (7):

[6] Emme myöskään sisällytä satunnaiskerroinmalliin murrealuetta. Tämä johtuu siitä, että satunnaisvakion ja satunnaiskertoimen välinen korrelaatio oli liian korkea murrealueen osalta (0.89), mikä viittaa siihen, että malli on liian monimutkainen.

[7] Lisäksi kuvioiden tekemiseen käytimme paketteja `lattice` (Sarkar 2008), `ggplot2` (Wickham 2016) ja `sjplot` (Lüdecke 2020).

(7) `kongruenssi ~ 1 + syntymäaika + sukupuoli +`
`dependenssin_pituus + pronomini +`
`(1|verbi_lemma) + (1|murrealue/murreryhmä/haastateltava)`

Notaatiossa ”kongruenssi” tarkoittaa selitettävää tekijää ja aaltoviivaa (~) seuraavat termit ovat selittäviä tekijöitä (`1` tarkoittaa vakiotermiä). Kolmannella rivillä on satunnaistermit. Notaatio `(1|verbi_lemma)` tarkoittaa, että mallin vakiotermin annetaan vaihdella verbin lemman mukaan. Notaatio `(1|murrealue/murreryhmä/haastateltava)` tarkoittaa, että mallin vakiotermin annetaan lisäksi vaihdella haastateltavan mukaan ottaen kuitenkin samalla huomioon, että kukin haastateltava edustaa tiettyä murreryhmää ja tietty murreryhmä kuuluu aina samaan murrealueeseen (esim. Pohjois-Savon murreryhmä kuuluu savolaismurteiden murrealueeseen; ks. taulukko 1). Satunnaiskerroinmallille käytimme `lme4`:n notaatiota (8):

(8) `kongruenssi ~ 1 + syntymäaika + sukupuoli +`
`dependenssin_pituus + pronomini +`
`(1|verbi_lemma) + (1+pronomini|murreryhmä/haastateltava)`

Yleistetyssä lineaarisessa sekamallissa yksi yleinen ongelma etenkin satunnaiskerroinmallien kohdalla on mallien konvergoituminen. Mallin parametrit lasketaan iteratiivisen algoritmin avulla, ja tämä prosessi pysähtyy, kun ero edelliseen iteraatioon on raja-arvoa pienempi. Yksi ratkaisu tähän ongelmaan on optimoida algoritmi vakioasetuksista poikkeavalla tavalla, käyttämällä esimerkiksi `BOBYQA`-optimoijaa kaikille satunnaistermeille. Käytimme tätä keinoa kaikissa malleissamme, koska osa niistä ei konvergoitunut algoritmin vakioasetuksilla.

Rakensimme mallimme askel askeleelta lähtien ns. nollamallista, joka sisältää ainoastaan vakiotermin. Tähän malliin lisäsimme asteittain ensin satunnaistermit ja sitten selittävät termit. Jokaisella askeleella vertasimme toisiinsa mallia, joka sisälsi yhden tekijän enemmän kuin toinen malli, ja samalla vertasimme toisiinsa mallien Akaiken informaatiokriteerin arvoja (engl. AIC = Akaike Information Criterion). AIC kuvaa mallin kompleksisuuden ja sen tarkkuuden välistä suhdetta, ja se on sitä pienempi, mitä parempi malli on. Yleensä malli on syytä pitää mahdollisimman yksinkertaisena kuitenkaan poistamatta siitä sellaisia tekijöitä, joiden takia menetettäisiin liikaa informaatiota. Näin ollen mitä enemmän jokin tekijä vähentää AIC:n arvoa, sitä tärkeämpi se on sisällyttää malliin. Burnham ja Anderson (2002: 70–71) esittävät karkeat suuntaviivat sille, miten tulkita tekijän tärkeys AIC:n valossa. Jos tietyn tekijän lisääminen malliin pienentää AIC:n arvoa vähemmän kuin 2 yksikköä, verrattavien mallien välillä ei juuri ole eroa. Tällöin lisättävä tekijä on turha eikä sitä tarvita lopullisessa mallissa. Jos taas tietyn tekijän lisääminen malliin pienentää AIC:n arvoa 4–7 yksikköä, tällöin voidaan sanoa, että malli, jolla on pienempi AIC, on parempi. Tällaisissa tilanteissa lisättävä tekijä on syytä pitää mukana mallissa. Jos taas tietyn tekijän lisääminen malliin pienentää AIC:n arvoa enemmän kuin 10 yksikköä, tällöin sille mallille, jolla on korkeampi AIC, ei esiinny lainkaan tilastollista tukea. Toisin sanoen lisättävä tekijä on ehdottomasti pidettävä

mukana mallissa. Tämän valintaprosessin lopuksi päädytään ns. parhaaseen malliin, josta on karsittu pois sellaiset tekijät, joita ei ole perusteltua sisällyttää malliin.

3 Tulokset

Esitämme ensin tulokset satunnaisvakiomallista, jolla tarkastelemme lukukongruenssin vaihtelua koko aineistossa käyttäen rakenteeltaan yksinkertaisinta monitasomallia. Tämän jälkeen esitämme satunnaiskerroinmallin tulokset. Aloitamme rakentamalla mallin askel askeleelta lisäämällä aina yhden termin ja vertaamalla malleja toisiinsa Akaiken informaatiokriteerin (AIC) avulla, minkä jälkeen tarkastelemme parhaan mallin testisuureita.

Lähdemme liikkeelle ns. nollamallista, joka sisältää ainoana selittävänä tekijänä vakiotermin. Sitten rakennamme mallia asteittain lisäämällä ensin satunnaistermit ja sen jälkeen selittävät tekijät, kunnes kaikki mallin termit on lisätty. Tätä prosessia kuvaamme taulukon 2 ensimmäisessä sarakkeessa. Otamme jokaisesta mallista AIC:n arvon, joka on esitetty taulukon 2 toisessa sarakkeessa. Taulukon kolmas sarake esittää AIC-arvojen erotuksen tietylle mallille ja sitä asteen yksinkertaisemmalle mallille. Esimerkiksi mallien 2 ja 1 AIC-arvojen erotus on 424,8. Tämä tarkoittaa sitä, että satunnaistermin ”verbin lemma” osuus mallin 2 AIC-arvosta on 424,8. AIC-arvojen erotusten perusteella voidaan sanoa, että mallimme kaikki satunnaistermit ovat erittäin tärkeitä, koska kukin niistä pienentää AIC-arvoa yli 400 yksikköä. Puhujan sukupuoli parantaa mallia vain rahtusen (0.3), joten se on syytä jättää pois tästä mallista. Puhujan syntymäaika huonontaa mallia, koska se kasvattaa eikä vähennä AIC-arvoa. Dependenssin pituus pienentää mallin AIC arvoa 153,7 yksikön verran ja subjektipronomini 18,8 yksikön verran, joten kumpikin niistä parantaa mallia selvästi.

Taulukko 2. Satunnaisvakiomallin rakentaminen asteittain ja mallien vertailu Akaiken informaatiokriteerin avulla (AIC:n erotus = kahden mallin AIC-arvojen erotus)

Mallin numero	Mallin rakenne	AIC	AIC:n erotus	Residuaalin vapausasteet
1	Vakiotermi	9522,0		8960
2	+ satunnaistermi(verbin_lemma)	9097,2	-424,8	8959
3	+ satunnaistermi(murrealue)	7453,1	-1644,1	8958
4	+ satunnaistermi(murrealue:murreryhmä)	6491,3	-961,8	8957
5	+ satunnaistermi(murrealue:murreryhmä:puhuja)	5809,9	-681,5	8956
6	+ sukupuoli	5809,5	-0,3	8955
7	+ syntymäaika	5811,4	1,9	8954
8	+ dependenssin pituus	5657,7	-153,7	8953
9	+ pronomini	5638,9	-18,8	8952

AIC-arvojen mukaan parhaaseen malliin kuuluvat kaikki satunnaistermit sekä dependenssin pituus ja pronomini. Tämän mallin selittävien tekijöiden testisuureet on esitetty taulukossa 3 ja sen satunnaistermien varianssit ja niiden keskihajonnat on esitetty taulukossa 4. Dependenssin

pituus (0,61±0,05; $p < 0,0001$) ja pronomini (0,75±0,16; $p < 0,0001$) vaikuttavat kongruenssiin tilastollisesti erittäin merkitsevällä tavalla. Ilmikongruenssin todennäköisyys kasvaa noin 1,8-kertaiseksi ($e^{0,61}$) inkongruenssiin nähden aina, kun dependenssin pituus kasvaa yhdellä sanalla. Kun siis subjektin ja verbin välissä esiintyy kolme sanaa, ilmikongruenssi on tällöin jo 5,5 kertaa todennäköisempää kuin inkongruenssi. Ilmikongruenssi on lisäksi noin 2,1 kertaa ($e^{0,75}$) todennäköisempää pronominin *he* kanssa verrattuna pronominiin *ne*.

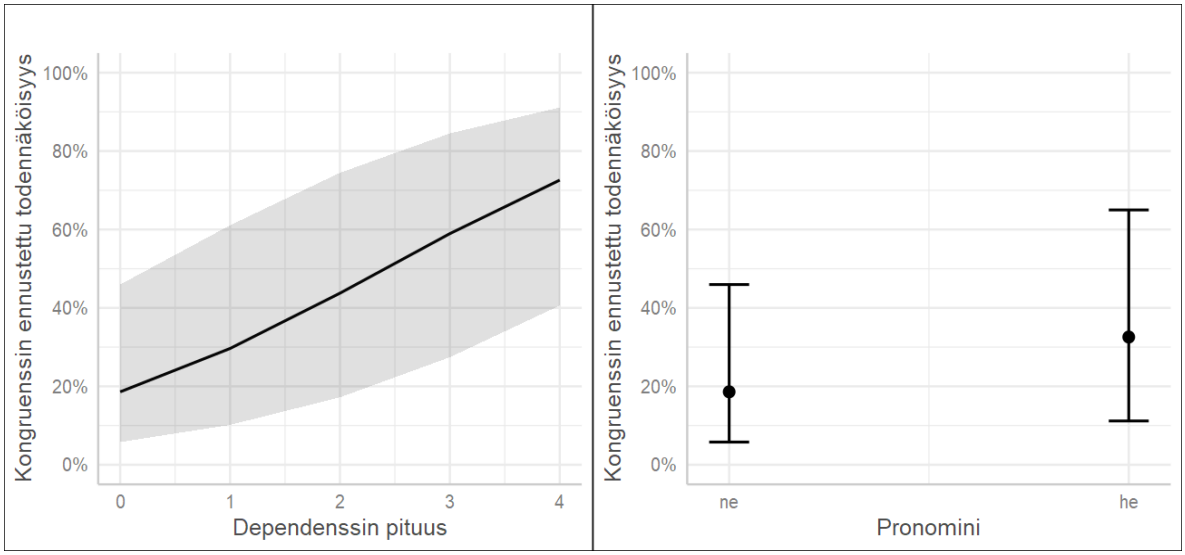
Taulukko 3. Parhaan satunnaisvakiomallin selittävät tekijät testisuureineen

	Estimaatti	Keskivirhe	z-arvo	p-arvo
(vakiotermi)	-1,4765	0,6707	-2,201	0,0277
dependenssin pituus	0,6127	0,0489	12,5400	< 0,0001
pronomini:he	0,7487	0,1606	4,6610	< 0,0001

Taulukko 4. Parhaan satunnaisvakiomallin satunnaistermit

Ryhmä	Termi	Varianssi	Keskihajonta
verbi_lemma	(vakiotermi)	0,4119	0,6418
haastateltava:(murreryhmä:murrealue)	(vakiotermi)	1,4278	1,1949
murreryhmä:murrealue	(vakiotermi)	2,2654	1,5051
murrealue	(vakiotermi)	1,8642	1,3654

Kuvio 3 esittää sovitetut arvot selittäville tekijöille ”dependenssin pituus” (vasemmalla) ja ”subjektipronomini” (oikealla). Kuviossa selittävien tekijöiden arvot ovat x-akselilla ja selitettävän tekijän eli kongruenssin sovitetut arvot ovat y-akselilla. Sovitetuista arvoista nähdään, että ilmikongruenssin todennäköisyys on noin 20 %, kun subjekti ja verbi ovat vierekkäin, mutta se kasvaa melko tasaisesti niin, että kun niiden välissä on neljä sanaa, ilmikongruenssin ennustettu todennäköisyys on jo yli 70 %. Lisäksi nähdään, että ilmikongruenssin todennäköisyys on noin 20 %, kun subjektina on *ne*-pronomini, mutta sen todennäköisyys nousee yli 30 %:iin, kun subjektina on *he*-pronomini.



Kuvio 3. Satunnaisvakiomallin sovitetut arvot dependenssin pituudelle ja pronominille

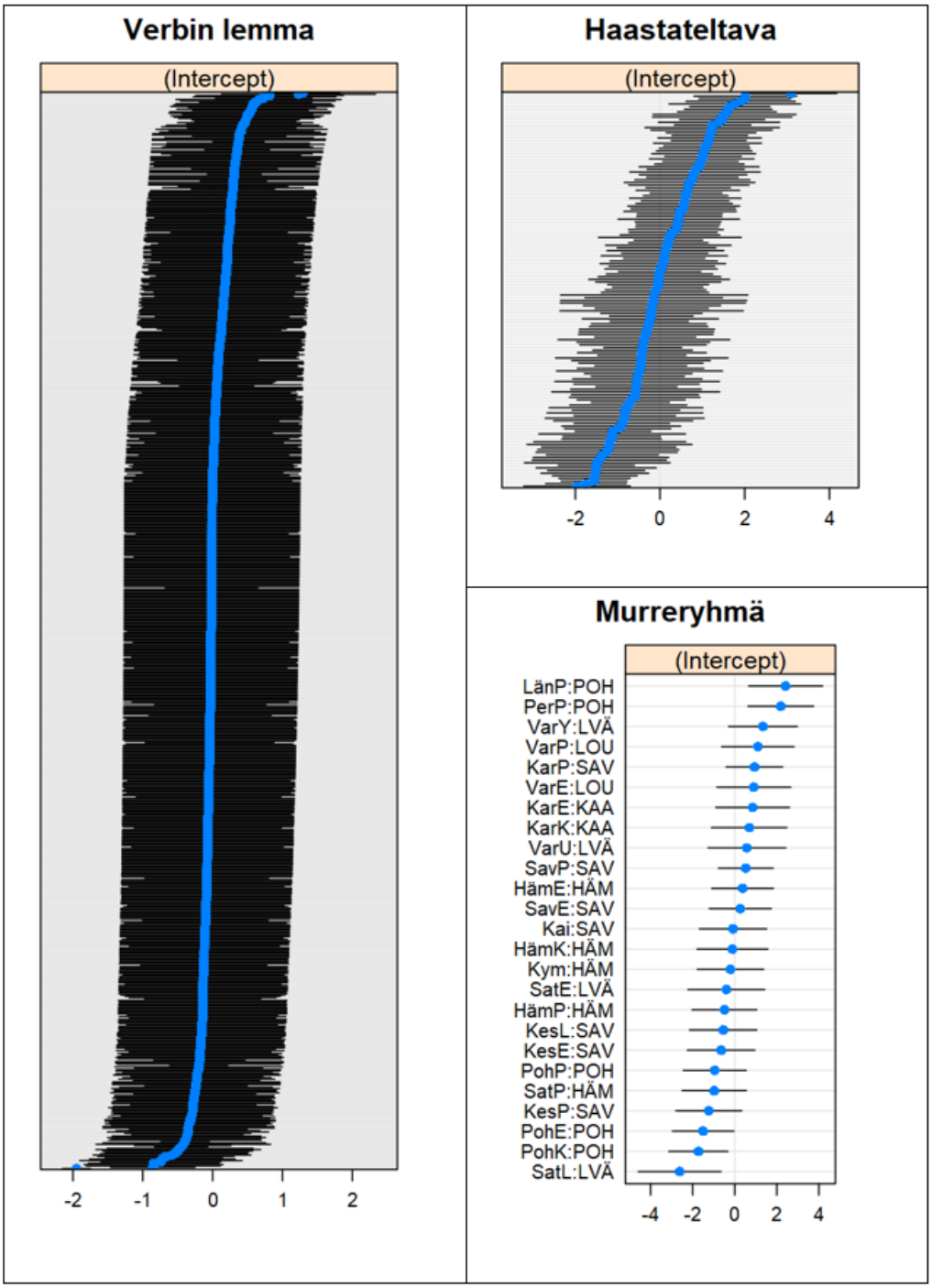
Regressiomallien yhteensopivuutta aineiston kanssa kuvataan usein selitysasteen R^2 avulla. Se kertoo, miten monta prosenttia selitettävän muuttujan varianssista selittävät muuttujat voivat selittää. Yleistetyissä lineaarisissa sekamalleissa käytetään selitysastetta R^2_{GLMM}, jonka avulla voidaan erottaa toisistaan kiinteiden termien ja satunnaistermien osuudet selitysasteesta.[8] Parhaan satunnaisvakiomallin kokonaisselitysaste on 59,0 %, mitä voidaan pitää hyvänä korpustutkimuksessa. Mallin selittävät tekijät selittävät kongruenssin vaihtelusta kuitenkin vain 1,8 %, kun taas satunnaistermit selittävät siitä 57,1 %. Suurin osa mallin selitysvoimasta johtuu siis verbin lemmasta sekä haastateltavasta ja hänen murteestaan. Vaikka rakenteellisten tekijöiden selitysvoima onkin vähäinen, mallimme onnistuu kuitenkin tunnistamaan niiden vaikutuksen puhujasta ja hänen murretaustastaan johtuvasta suuresta vaihtelusta huolimatta.

Kuvio 4 esittää kongruenssin vaihtelun eri satunnaistermien suhteen. Kuviossa vasemmalla on verbin lemmasta johtuva vaihtelu, oikealla ylhäällä haastateltavasta johtuva vaihtelu (kiinnitettynä murreryhmään) ja oikealla alhaalla murreryhmästä johtuva vaihtelu (kiinnitettynä murrealueeseen). Yksittäisen verbin lemman vaikutus kongruenssiin poikkeaa keskimäärin vain vähän nollasta. Suurimmat positiiviset poikkeamat koskevat verbejä *viedä* (1,29) ja *kierrellä* (1,23), joilla monikkokongruenssia esiintyy keskimääräistä selvästi enemmän verrattuna muihin verbeihin. Suurin negatiivinen poikkeama keskiarvosta esiintyy *olla*-verbillä (–1,95), joka esiintyy murteissa tyypillisesti yksikkömuodossa myös monikollisten subjektipronominien kanssa. Yksittäisen haastateltavan vaikutus kongruenssiin poikkeaa keskimäärin myös melko vähän nollasta.

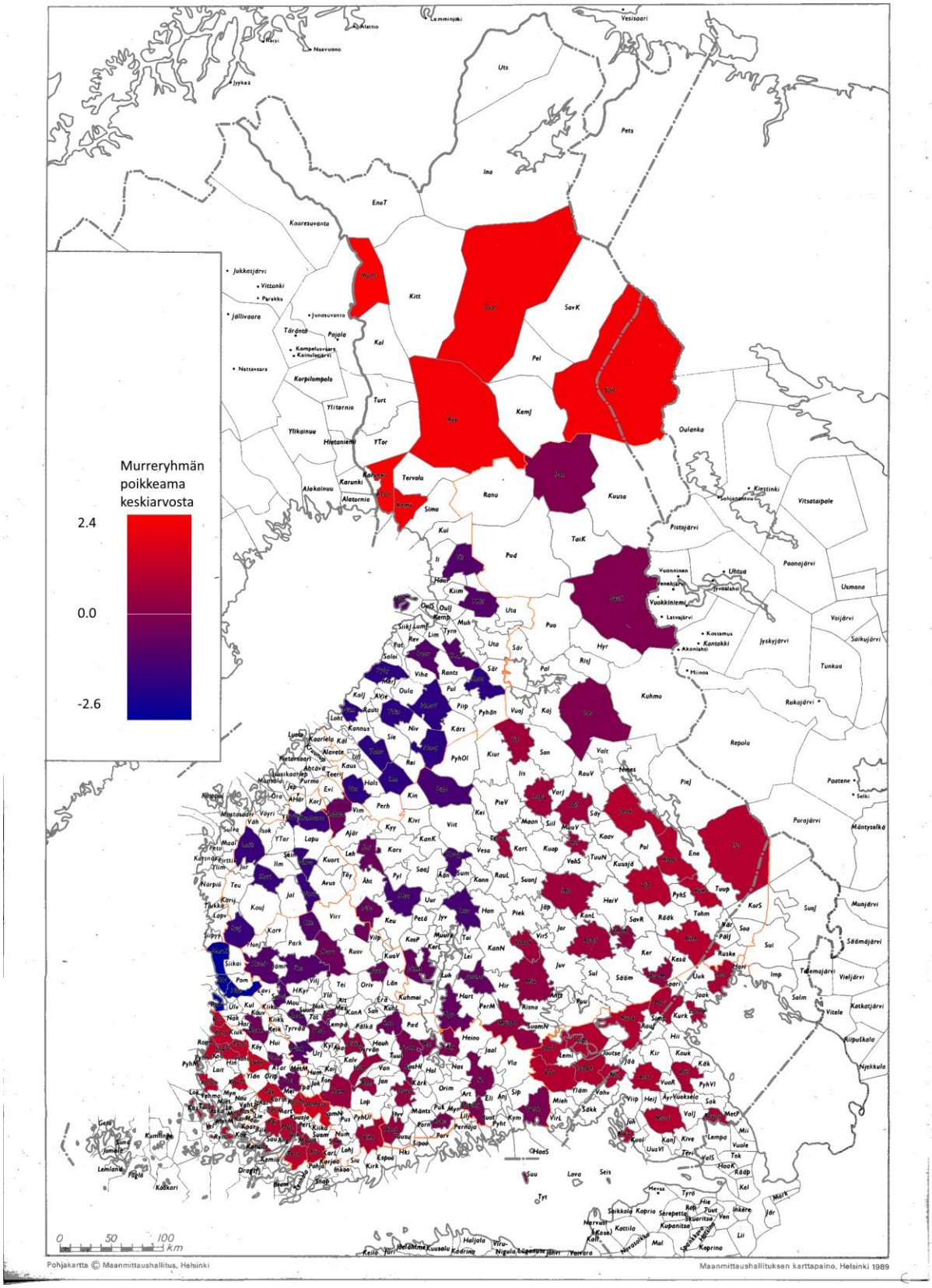
Kongruenssi vaihtelee kiinnostavalla tavalla eri murreryhmissä. Suurimmat negatiiviset poikkeamat keskiarvosta (< –1) esiintyvät Länsi-Satakunnan (SatL), Keski-Pohjanmaan (PohK), Etelä-Pohjanmaan (PohE) ja Pohjoisen Keski-Suomen (KesP) murreryhmissä. Näissä murreryhmissä monikkokongruenssia esiintyy siis keskimääräistä vähemmän. Suurehkoja

[8] Testisuure lasketaan R-paketin `MuMIn` avulla (Barton 2019). Algoritmi perustuu Nakagawan ja Schielzethn (2013) työhön, jota ovat jatkokehitelleet Johnson (2014) sekä Nakagawa, Johnson ja Schielzeth (2017).

negatiivisia poikkeamia (< –0,5) esiintyy myös Pohjois-Satakunnan (SatP), Pohjois-Pohjanmaan (PohP) ja Pohjois-Hämeen (HämP) murreryhmissä sekä Eteläisen Keski-Suomen (KesE) ja Läntisen Keski-Suomen (KesL) murreryhmissä. Suurehkoja positiivisia poikkeamia keskiarvosta (> 0,5) esiintyy Länsipohjan ja Peräpohjolan murreryhmissä, Eteläisen Varsinais-Suomen (VarE) ja Pohjoisen Varsinais-Suomen (VarP) murreryhmissä (so. lounaismurteissa), ja Etelä-Karjalan (KarE) ja Keski-Karjalan murreryhmissä (KarK) (so. kaakkoismurteissa) sekä lisäksi Pohjois-Karjalan (KarP) ja Varsinais-Suomen ylämaan (VarY) murreryhmissä.



Kuvio 4. Kongruenssin vaihtelu parhaassa satunnaisvakiomallissa eri satunnaisvakioilla. Vasemmalla kongruenssin vaihtelu verbin lemman suhteen, oikealla ylhäällä haastateltavan suhteen ja oikealla alhaalla murreryhmän suhteen



Kuvio 5. Kartta monikkokongruenssin vaihtelusta murreryhmittäin; arvot ovat poikkeamia keskiarvosta. Karttapohja on Maanmittauslaitoksen tekemä, mutta se on vapaasti käytettävissä

Kuvio 5 esittää kongruenssin vaihtelun kartalla murreryhmittäin. Tiettyyn murreryhmään kuuluva pitäjä on sävytetty sitä syvemmän sinisellä, mitä vähemmän lukukongruenssia murreryhmässä esiintyy verrattuna keskimääräiseen ja sitä kirkkaamman punaisella, mitä enemmän lukukongruenssia murreryhmässä esiintyy verrattuna keskimääräiseen. Arvot lämpökarttaan on otettu murreryhmän satunnaisvakioista kuviosta 4. Ne murreryhmät, joissa monikkokongruenssia esiintyy keskimääräistä vähemmän, näyttävät muodostavan yhtenäisen vyöhykkeen Oulusta Poriin ja se ulottuu paikoin syvälle sisämaahan. Niitä murrealueita, joissa monikkokongruenssia esiintyy keskimääräistä enemmän, ei kuitenkaan yhdistä toisiinsa mikään ilmeinen maantieteellinen seikka: ne ovat kaukana toisistaan ja niiden välissä esiintyy murreryhmiä, jotka poikkeavat vain vähän keskiarvosta tai poikkeavat siitä negatiivisesti. Lukukongruenssin alueellisen jakauman perusteella on mahdollista, että sen vähäisempi käyttö on innovaatio läntisissä murreryhmissä, joista se on alkanut levitä myös sisämaan murreryhmiin. Asiaa olisi syytä tutkia perusteellisemmin jatkotutkimuksessa.

Kaiken kaikkiaan satunnaisvakiomallissa rakenteelliset selittävät tekijät vaikuttavat kongruenssin vaihteluun, mutta sosiolingvistiset selittävät tekijät eivät. Verbin lemma ja yksittäisestä puhujasta ja hänen murretaustastaan johtuva vaihtelu oli erittäin tärkeä ottaa huomioon, koska ne selittivät suurimman osan kongruenssin vaihtelusta aineistossa.

Seuraavaksi tarkastelemme kompleksisempaa satunnaiskerroinmallia, jossa otamme huomioon sen, että pronominin vaikutus kongruenssiin voi vaihdella riippuen yksittäisestä puhujasta ja hänen murteestaan. Tämän mallin tarkastelussa käytämme vain osaa aineistosta (ks. luku 2.2). Aloitamme jälleen rakentamalla mallin askel askeleelta ja vertaamalla mallien AIC-arvoja toisiinsa (ks. taulukko 5). Tässä vertailussa otamme huomioon ainoastaan satunnaiskertoimen lisäämisen sekä sen vaikutuksen selittäviin tekijöihin.

AIC-arvojen erotusten perusteella satunnaiskertoimen lisääminen vähensi mallin AIC-arvoa 61,6 yksikköä. Tämän perusteella satunnaiskerroin on tärkeä mallissa. Selittävistä tekijöistä puhujan sukupuoli vähentää AIC-arvoa 2,0 yksikköä eli se parantaa mallia jonkin verran. Puhujan syntymäaika kuitenkin jälleen huonontaa mallia, koska sen lisääminen malliin kasvattaa eikä vähennä AIC-arvoa. Dependenssin pituus pienentää mallin AIC arvoa 78,5 yksikköä ja subjektipronomini 4,5 yksikköä, joten kumpikin niistä on syytä pitää mukana parhaassa mallissa puhujan sukupuolen lisäksi.

Taulukko 5. Satunnaiskerroinmallin rakentaminen asteittain ja AIC-arvojen vertailu

Mallin nro	Mallin rakenne	AIC	AIC:n erotus	Residuaalin vapausasteet
1	Vakiotermi + satunnaisvakio (1|verbi_lemma) + satunnaisvakio (1|murreryhmä/puhuja)	4076,7		6036
2	+ satunnaiskerroin (1+pronomini|murreryhmä/puhuja)	4015,1	-61,6	6032
3	+ sukupuoli	4013,2	-2,0	6031
4	+ syntymäaika	4015,2	2,0	6030
5	+ dependenssin pituus	3936,7	-78,5	6029
6	+ pronomini	3932,1	-4,5	6028

Taulukko 6. Parhaan satunnaiskerroinmallin selittävät tekijät testisuureineen

	Estimaatti	Keskivirhe	z-arvo	p-arvo
(vakiotermi)	-1,5247	0,7087	-2,151	0,0314
sukupuoli:mies	-0,6717	0,3328	-2,018	0,0436
dependenssin pituus	0,5359	0,0595	9,006	< 0,0001
pronomini:he	1,3866	0,4674	2,967	0,003

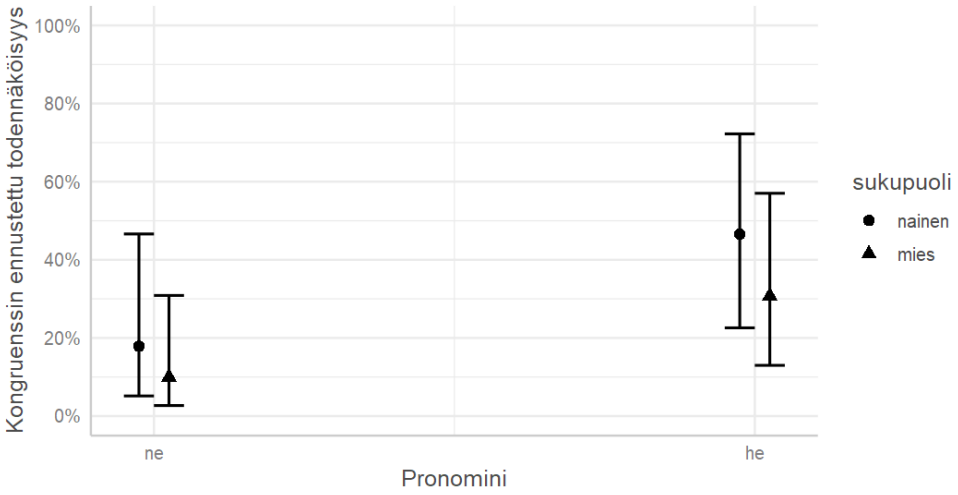
Taulukko 7. Parhaan satunnaiskerroinmallin satunnaistermit

Ryhmä	Termi	Varianssi	Keskihajonta	Korrelaatio
verbi_lemma	(vakiotermi)	0,4898	0,6999	
haastateltava:murreryhmä	(vakiotermi)	1,8093	1,3451	
	pronomini:he	1,1036	1,0505	-0,63
murreryhmä	(vakiotermi)	6,5803	2,5652	
	pronomini:he	2,2371	1,4957	-0,72

AIC-arvojen perusteella parhaaseen satunnaiskerroinmalliin kuuluvat kaikki satunnaistermit ja selittävinä tekijöinä dependenssin pituus, pronomini ja puhujan sukupuoli. Tämän mallin selittävien tekijöiden testisuureet on esitetty taulukossa 6 ja sen satunnaistermit taulukossa 7. Satunnaistermien korrelaatiot ovat edelleen korkeahkoja (–0,63 ja –0,72), mutta eivät silti liian korkeita. Kukin selittävä tekijä, dependenssin pituus (0,54±0,06; $p < 0,0001$), pronomini (1,39±0,47; $p = 0,003$) ja puhujan sukupuoli (–0,67±0,33; $p = 0,044$), vaikuttaa kongruenssiin tilastollisesti merkitsevällä tavalla. Ilmikongruenssin todennäköisyys kasvaa noin 1,7-kertaiseksi ($e^{0,54}$) inkongruenssiin nähden aina kun dependenssin pituus kasvaa yhdellä sanalla. Kun siis subjektin ja verbin välissä on kolme sanaa, ilmikongruenssi on tällöin jo 5,1 kertaa todennäköisempää kuin inkongruenssi. Ilmikongruenssi on lisäksi noin neljä kertaa ($e^{1,39}$)

todennäköisempää pronominin *he* kanssa verrattuna pronominiin *ne* ja noin kaksi kertaa vähemmän todennäköistä miesten puheessa verrattuna naisten puheeseen ($e^{-0,67}$).

Kuvio 6 esittää sovitetut arvot selittäville tekijöille ”subjektipronomini” ja ”sukupuoli” samassa kuviossa. Kuviossa pronominin arvot ovat x-akselilla ja kongruenssin sovitetut arvot ovat y-akselilla. Sovitetuista arvoista nähdään, että pronominin *ne* osalta ilmikongruenssin todennäköisyys on miehillä noin 10 % ja naisilla noin 20 %. Pronominin *he* osalta ilmikongruenssin todennäköisyys on miehillä noin 30 % ja naisilla lähes 50 %. Näiden arvojen perusteella lukukongruenssin todennäköisyys naisilla ja miehillä ei eroa kovin paljon *ne*-pronominin yhteydessä, mutta se eroaa selvästi *he*-pronominin yhteydessä.



Kuvio 6. Satunnaiskerroinmallin sovitetut arvot pronominille ja sukupuolelle

Tilastolliset tulokset antavat aihetta seuraaviin yleistyksiin. Dependenssin pituus vaikuttaa sekä satunnaiskerroinmallissa että satunnaisvakiomallissa lukukongruenssin vaihteluun samalla tavalla: mitä suurempi etäisyys subjektin ja verbin välillä esiintyy, sitä todennäköisemmin monikkokongruenssi esiintyy verbissä. Satunnaiskerroinmallissa otimme huomioon sen, että pronominin vaikutus lukukongruenssiin voi vaihdella riippuen yksittäisestä puhujasta ja hänen murteestaan. Tästä huolimatta pronomini vaikuttaa monikkokongruenssiin tilastollisesti merkitsevällä tavalla. Lisäksi puhujan sukupuoli vaikuttaa satunnaiskerroinmallissa monikkokongruenssiin siten, että se on todennäköisempää naisilla kuin miehillä ja etenkin *he*-pronominin yhteydessä.

4 Tulkinta

Tässä tutkimuksessa olemme selvittäneet differentiaalisen lukukongruenssin vaihtelua suomen murteissa Lauseopin arkiston murrekorpuksessa. Yleistetyn lineaarisen sekamallin mukaan lukukongruenssin vaihteluun vaikuttavat dependenssin pituus ja subjektipronomini sekä jossain määrin puhujan sukupuoli mutta ei ikä. Lisäksi siihen vaikuttavat verbin lemma sekä

yksittäinen puhuja ja hänen murteensa. Seuraavassa tulkitsemme näitä tuloksia mm. typologisen tutkimuksen näkökulmasta.

Tulosten perusteella differentiaalinen lukukongruenssi riippuu dependenssin pituudesta subjektipronominin ja predikaattiverbin välillä siten, että monikkokongruenssin todennäköisyys kasvaa, mitä etäämmällä subjekti esiintyy verbistä. Tällainen morfologisen ilmimerkinnän vaihtelu dependenssin pituuden suhteen vaikuttaa olevan universaali ilmiö maailman kielissä (Hawkins 2004). Mitä lähempänä dependentti ja pääsana ovat toisiaan, sitä helpompi kuulijan on päätellä dependenssisuhde ilman morfologista merkintää, ja mitä etäämmällä dependentti on pääsanasta, sitä vaikeampi kuulijan on prosessoida dependenssisuhde ilman suhteen morfologista merkitsemistä joko dependenttiin tai pääsanaan. Funktionaalisessa kielitieteessä morfologisen merkinnän todennäköisyyden lisääntyminen dependenssin pituuden kasvaessa liittyy ekonomiaan eli siihen, että puhujat pyrkivät ottamaan kuulijan viestinnälliset tarpeet huomioon mahdollisimman tehokkaasti (esim. Zipf 1949 ja vähimmän vaivan periaate; F. Karlsson 1977). Viime aikoina ekonomiaan liittyviä ajatuksia on systematisoitu yhden teoreettisen viitekehyksen alle (mm. Hawkins 2014, Gibson ym. 2019).

Malliemme mukaan lukukongruenssiin vaikuttaa kuitenkin myös subjektipronominin tyyppi: ilmikongruenssi on selvästi todennäköisempää, kun monikollisena subjektina toimii pronomini *he* verrattuna pronominiin *ne*. Tällainen kongruenssin systemaattinen vaihtelu nominityypin mukaan on määritelmällisesti esimerkki substantiiviluokasta (esim. Corbett 2000; Nichols 1992).[9] Vaikka vaihtelu aineistossa onkin systemaattista, se on tilastollista eikä kategorista, jollaista substantiiviluokka on esimerkiksi monissa bantukielissä: ilmikongruenssia esiintyy myös *ne*-pronominin yhteydessä ja inkongruenssia *he*-pronominin yhteydessä. Differentiaalista lukukongruenssia voisikin ehkä parhaiten kuvata Bresnania, Dingareeta ja Manningia (2001) mukaillen ns. pehmeänä rajoitteena. Sen mukaan tilastollinen ja poikkeuksia sisältävä trendi voi yhtä kaikki olla samansuuntainen tiukan kategorisen erottelun kanssa.

Artikkelin johdannossa viittasimme siihen, että substantiiviluokan kannalta ei ole oleellista se, onko kongruenssin vaihteluun liittyvä nominityyppien välinen ero semanttinen vai ei. Pronominien *he* ja *ne* välinen ero liittyy jossain määrin tarkoitteiden elollisuuteen, sillä *he*-pronominia käytetään useammin henkilötarkoitteiden kanssa. Kirjakielessä *he*-pronominin käyttö on rajattu nimenomaan viittaamaan henkilöihin, kun taas *ne*-pronominin käyttö on rajattu viittaamaan muihin tarkoitteisiin. Murteissa pronominilla *he* on lisäksi logoforisuuteen liittyviä funktioita, vastaavasti kuin pronominilla *hän*, mutta ne eivät kuitenkaan ole varsinaisia logoforisia pronomineja (pronominin *hän* osalta ks. Priiki 2017: 25–27). Pronominilla *ne* on lisäksi impersonaalisuuteen liittyviä funktioita (Forsberg ym. 2009).

Tässä tutkimuksessa ei ollut mahdollista analysoida kaikkia subjektipronominien tarkoitteita elollisuuden suhteen. Otimme kuitenkin noin 130 lauseen satunnaisotoksen

[9] Kutsumme tätä ilmiötä mieluiten substantiiviluokaksi, vaikka esim. Corbett (1991) ja Nichols (1992) kutsuvat sitä kieliopillisen suvun merkinnäksi. Miellämme suvun merkinnän substantiiviluokan erikoistyypiksi, johon liittyy tai on liittynyt biologisen sukupuolen mukainen semanttinen ero.

Lauseopin arkiston murrekorpuksesta ja analysoimme pronominien *ne* ja *he* tarkoitteet niiden elollisuuden mukaan (eloton, elollinen, ihmistarkoite). Kuvio 7 esittää yhteenvedon analyysista. Pilottitutkimuksen perusteella pronomini *ne* (n = 32/96; 33 %) viittaa selvästi useammin elottomiin tarkoitteisiin verrattuna pronominiin *he* (n = 1/36; 2 %). Silti pääosa *ne*-pronominin tarkoitteista on elollisia tai ihmistarkoitteisia (n = 64/96; 67 %). Pronominin *he* tarkoitteista lähes kaikki ovat ihmistarkoitteisia (n = 35/36; 87,5 %).

Kuvio 7. Verbin lukukongruenssi ja subjektipronominien tarkoitteiden elollisuus

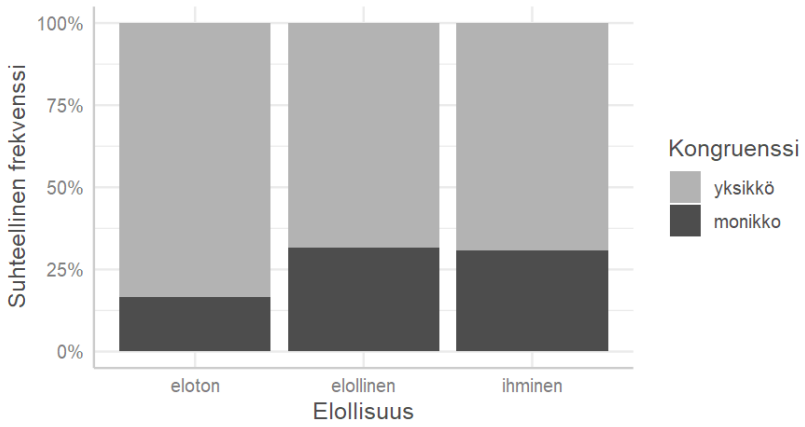
Kuvion 7 perusteella noin puolessa pronominin *he* ihmisviitteisistä tapauksista verbissä on luvun ilmikongruenssi, kun taas pronominin *ne* tapauksista luvun ilmikongruenssia esiintyy noin 17–26 %:ssa tapauksista riippuen tarkoitteen elollisuudesta. Ihmistarkoitteisissa tapauksissa ilmikongruenssi onkin noin kolme kertaa todennäköisempää pronominin *he* yhteydessä verrattuna pronominiin *ne* (p = 0,015; Fisherin tarkka testi). Eroa ei kuitenkaan esiinny ei-ihmistarkoitteisissa tapauksissa (p = 0,57), vaikka ilmikongruenssia esiintyykin niissä enemmän pronominin *ne* yhteydessä. Jakaumien perusteella voidaan alustavasti todeta, että pronominien *he* ja *ne* välinen ero liittyy ainakin jossain määrin tarkoitteen elollisuuteen ja tuo vaihtelu liittyy myös lukukongruenssin vaihteluun.

Rajasimme tässä artikkelissa differentiaalisen lukukongruenssin tutkimuksen sellaisiin tapauksiin, joissa subjektina esiintyy monikollinen subjektipronomini *he* tai *ne*. Voidaan kuitenkin perustellusti kysyä, liittyykö aineistossa havaitsemamme evidenssi nominityypin ja differentiaalisen lukukongruenssin välillä pelkästään subjektipronomineihin. Työmäärän rajoittamiseksi emme sisällyttäneet tähän tutkimukseen kattavaa analyysia leksikaalisista subjekteista. On kuitenkin oletusarvoista, että differentiaalinen lukukongruenssi suomen murteissa liittyy yleisemminkin subjektin elollisuuteen eikä pelkästään pronominien *he* ja *ne* vaihteluun. Tätä selvittääksemme teimme pienen pilottianalyysin, jossa poimimme Lauseopin arkiston murrekorpuksesta satunnaisesti noin 200 lausetta, joissa verbiä edelsi välittömästi

monikkomuotoinen leksikaalinen subjekti. Analysoimme kunkin subjektin elollisuuden elollisuushierarkian (1) mukaisesti, jonka toistamme sovellettuna hierarkiassa (9).

(9) Elollisuushierarkia (Corbett 2010: 92)
ihminen > muut elolliset > eloton

Kuvio 8 esittää lukukongruenssin jakauman elollisuuden suhteen otoksen leksikaalisilla monikkosubjekteilla. Ilmikongruenssin frekvenssi on samansuuntainen ihmistarkoitteisissa ja muissa elollisissa leksikaalisissa monikkosubjekteissa. Lisäksi ilmikongruenssin frekvenssi on kummassakin luokassa selvästi suurempi kuin elottomissa monikkosubjekteissa. Kun ryhmittelemme ihmistarkoitteiset ja muut elolliset leksikaaliset monikkosubjektit yhdeksi luokaksi, ilmikongruenssi on niissä noin 2,3 kertaa todennäköisempää kuin elottomissa leksikaalisissa monikkosubjekteissa ($p = 0{,}028$; Fisherin tarkka testi).



Kuvio 8. Verbin lukukongruenssi ja leksikaalisten monikkosubjektien tarkoitteiden elollisuus

Tämän rajallisen pilottitutkimuksen perusteella näyttää siis siltä, että differentiaalinen lukukongruenssi riippuu tilastollisesti merkitsevällä tavalla leksikaalisen subjektin elollisuudesta. Tulos antaa viitteitä siitä, että differentiaalinen lukukongruenssi suomen murteissa saattaa olla yleisemminkin suhteessa subjektin elollisuuteen eikä rajoitu pelkästään pronomineihin *he* ja *ne*. Jatkotutkimus asiasta on vireillä, jotta ilmiöstä saataisiin tarkempi kuva.

Differentiaaliseen lukukongruenssiin näyttää Lauseopin arkiston murrekorpuksen valossa vaikuttavan eniten puhujasta ja hänen murteestaan johtuva vaihtelu. Tähän viittasivat sekä satunnaisvakiomallin selitysaste että mallien AIC-arvojen vertailu. Aiemmassa tutkimuksessa murteiden välillä onkin havaittu selviä eroja lukukongruenssin vaihtelussa (mm. G. Karlsson 1966). Tilastolliset tuloksemme vahvistavat näitä aiempia tuloksia. Murrevaihtelun osalta on kiinnostavaa, että murreryhmät, joissa ilmikongruenssin todennäköisyys on keskimääräistä pienempi, muodostavat yhtenäisen maantieteellisen alueen, joka ulottuu suurin piirtein Oulusta Poriin ja syvällekin sisämaahan. Sitä vastoin murreryhmät, joissa ilmikongruenssin todennäköisyys on keskimääräistä suurempi, eivät muodosta ainakaan maantieteellisesti

yhtenäistä ryhmää. On kuitenkin epäselvää, miksi juuri läntisissä murreryhmissä Oulusta Poriin monikkokongruenssi on keskimääräistä vähäisempää. Inkongruenssin taajempi käyttö on saattanut olla innovaatio tällä alueella, josta se on levinnyt myös muualle. Aihe vaatii kuitenkin jatkotutkimusta etenkin historiallisen murretutkimuksen näkökulmasta.

Lukukongruenssiin vaikuttaa tuloksen perusteella myös verbin lemma. Etenkin *olla*-verbillä ilmikongruenssi on selvästi vähäisempää verrattuna muihin verbeihin. Ilmiö on tuttu jo aiemmasta tutkimuksesta (G. Karlsson 1966; Mielikäinen 1984: 167), mutta mitään selvää syytä ei ole esitetty kopulaverbin poikkeavalle käyttäytymiselle. Yksi mahdollinen syy saattaa liittyä verbin transitiivisuuteen ja verbin argumenttien erotteluun. Mitä enemmän verbillä on argumentteja, sitä enemmän kieliopillisia vinkkejä tyypillisesti tarvitaan argumenttien erotteluun, toisin sanoen joko useita eri sijoja tai yhdessä sijaa, kongruenssia ja sanajärjestystä. Kopulaverbillä tällaista tarvetta on kaikkein vähiten. Tämän artikkelin puitteissa tätä hypoteesia ei ole mahdollista testata, mutta aihe on kiinnostava jatkotutkimuksen kannalta.

Analyysimme perusteella aineistossa esiintyy tilastollista tukea sille, että puhujan sukupuoli vaikuttaa lukukongruenssiin. Kuten artikkelin johdannossa mainitsimme, Lauseopin arkiston murrekorpus edustaa tarkoituksella melko homogeenista joukkoa, jotta havaitut erot heijastaisivat nimenomaan murre-eroja. Korpuksen koostamisessa käytettyjen kriteerien vuoksi on yllättävää mutta varsin kiinnostavaa, että tulosten perusteella monikkokongruenssi näyttää olevan yleisempää naisten murrenäytteissä verrattuna miehiin. Tulosten perusteella tämä koskee etenkin *he*-pronominia, mutta tulos on samansuuntainen, joskaan ei yhtä vahva, *ne*-pronominin kohdalla. Tämä saattaa viitata siihen, että haastateltujen naisten puheessa on hienoista normitetun yleiskielen vaikutusta. Aihe vaatii kuitenkin jatkotutkimusta. Tulosten mukaan puhujan ikä ei vaikuta kongruenssin vaihteluun. Tämä on oletettavaa aineiston homogeenisuuden perusteella, koska suurin osa puhujista on syntynyt 20 vuoden sisällä.

5 Lopuksi

Tässä tutkimuksessa käytimme monimuuttujamenetelmiä differentiaalisen lukukongruenssin vaihtelun mallintamiseen. Nämä menetelmät ovat yleistyneet korpustutkimuksessa ja kokeellisessa tutkimuksessa, koska niiden avulla voidaan samanaikaisesti ottaa huomioon sekä selittävistä tekijöistä että mm. puhujasta ja lemmasta johtuva vaihtelu. Perinteisemmät monimuuttujamenetelmät, kuten logistinen regressio, eivät mahdollista kaikkien tällaisten muuttujien rakentamista malliin samanaikaisesti. Tilastollisen mallin avulla oli mahdollista vertailla eri tekijöiden vaikutusta lukukongruenssiin ja havaita myös heikompia vaikutuksia, jotka voi olla mahdotonta havaita suuresta aineistosta ilman laskennallisia menetelmiä.

Onkin huomattavaa, että riippumatta yksittäisen puhujan ja hänen murteensa suuresta vaikutuksesta lukukongruenssin vaihteluun, osa kongruenssin vaihtelusta on rakenteellisesti systemaattista ja johtuu dependenssin pituudesta ja nominisubjektin tyypistä. Maailman kielissä on tavallista, että kielellistä variaatiota rekrytoidaan erilaisiin funktioihin. Aineistomme perusteella differentiaalinen lukukongruenssi suomen murteissa on jossain määrin rekrytoitu substantiiviluokan ilmaisuun. Ilmiö on kiinnostava paljon suomen kielen ja

sen murteiden tutkimusta laajemmin, sillä se antaa potentiaalisesti aivan uutta tietoa substantiiviluokan kehittymisestä kieleen. Kaiken kaikkiaan tutkimuksemme tulokset vahvistavat useita aiemmassa tutkimuksessa tehtyjä havaintoja ja ounasteluja monikkokongruenssiin vaikuttavista tekijöistä. Näitä ovat mm. murteiden välinen vaihtelu (G. Karlsson 1966), dependenssin pituus (Mielikäinen 1984) ja monikollisen subjektipronominin valinta (Priiki 2017). Tutkimuksemme selkein uusi tulos liittyy juuri pronominin vaikutukseen lukukongruenssiin murteissa, mistä on vain vähän aikaisempia ounastelujakaan.

Termiä substantiiviluokka käytetään usein kielissä, joissa substantiivit jakaantuvat useaan, jopa yli kymmeneen eri luokkaan niiden kongruenssiparadigmojen mukaan. Näissä systeemeissä eri luokkien kesken ei välttämättä esiinny selviä semanttisia eroja, kuten elollisten tarkoitteiden sukupuoli tai elottomien tarkoitteiden koko, muoto tai tekstuuri. Tällaisia systeemejä esiintyy etenkin bantukielissä ja joissain Papua Uuden Guinean ja Amazonin kielissä. Termiä kieliopillinen suku tavataan käyttää sellaisista kielistä, joissa substantiivit jakaantuvat 2–4 kongruenssiparadigmaan ja joista ainakin historiallisesti yksi keskeisistä semanttisista tekijöistä liittyy elollisten tarkoitteiden sukupuoleen, kuten useissa indoeurooppalaisissa kielissä. On mahdollista, että tämä terminologinen perinne vaikuttaa siihen, mitä tutkijoina ylipäätään havainnoimme kielestä. Jos ilmiötä kutsutaan kieliopilliseksi suvuksi, se voi estää tutkijaa kiinnittämästä tarpeeksi huomiota potentiaalisesti differentiaaliseen kongruenssiin, joka liittyy eri nominityyppeihin, joskaan ei semanttisesti tarkoitteen sukupuoleen. Olisikin parempi nimittää tätä ilmiötä jatkossa substantiiviluokaksi ja kieliopillista sukua sen alaluokaksi.

Lyhenteet

1	ensimmäinen persoona
2	toinen persoona
3	kolmas persoona
ART	artikkeli
PL	monikko
REAL	reaalinen modus
PRF	perfektiivinen aspekti
SG	yksikkö

Lähteet

Aineistolähteet

Turun yliopisto, kieli- ja käännöstieteiden laitos, & Kotimaisten kielten keskus 1985: *Lauseopin arkiston murrekorpuksen Helsinki-Korp-versio* [tekstikorpus]. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2016040702.

Kirjallisuuslähteet

Baayen, R. Harald 2013: Multivariate statistics. Teoksessa Robert J. Podesva & Devyani Sharma (toim.) *Research Methods in Linguistics*, 337–372. Cambridge: Cambridge University Press.

Baayen, R. Harald, Davidson, Doug J. & Bates, Douglas M. 2008: Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59 (4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005.

Bamyacı, Elif, Häussler, Jana & Kabak, Barış 2014: The interaction of animacy and number agreement: An experimental investigation. *Lingua* 148, 254–277. https://doi.org/10.1016/j.lingua.2014.06.005.

Bayanati, Shiva, & Toivonen, Ida 2019: Humans, animals, things and animacy. *Open Linguistics* 5 (1), 156–170. https://doi.org/10.1515/opli-2019-0010.

Barton, Kamil 2019: MuMIn: Multi-Model Inference. R package version 1.43.6. https://CRAN.R-project.org/package=MuMIn [20.8.2020].

Bates, Douglas, Maechler, Martin, Bolker, Ben & Walker, Steve 2015: Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67 (1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Berg, René van den 1989: *A Grammar of the Muna Language*. Julkaistu elektronisena vuonna 2013 sarjassa SIL e-Books. Dordrecht: Foris. https://www.sil.org/resources/publications/entry/52170 [12.1.2021].

Bresnan, Joan, Dingare, Shipra & Manning, Christopher D. 2001: Soft constraints mirror hard constraints: Voice and person in English and Lummi. Teoksessa *Proceedings of the LFG01 Conference*, 13–32. Stanford: CSLI Publications.

Burnham, Kenneth P. & Anderson, David R. 2002: *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2. laitos. New York: Springer.

Corbett, Greville G. 1991: *Gender*. Cambridge: Cambridge University Press.

Corbett, Greville G. 2000: *Number*. Cambridge: Cambridge University Press.

Corbett, Greville G. 2010: *Features*. Cambridge: Cambridge University Press.

Dahl, Östen 2000: Elementary gender distinctions. Teoksessa Matti Rissanen, Terttu Nevalainen & Mirja Saari (toim.) *Gender in Grammar and Cognition. Vol. 2: Elementary Gender Distinctions: Manifestations of Gender*, 577–593. Berliini: Mouton de Gruyter.

De Swart, Peter 2007: *Cross-linguistic Variation in Object Marking*. Väitöskirja, Radbound University Nijmegen. https://www.lotpublications.nl/Documents/168_fulltext.pdf [12.1.2021].

Di Garbo, Francesca, Olsson, Bruno & Wälchli, Bernhard (toim.) 2019: *Grammatical Gender and Linguistic Complexity*, 2 nidettä. Berliini: Language Science Press. https://doi.org/10.5281/zenodo.3446224.

Ellonen, Noora 2006: Monitasoanalyysit ja niiden soveltaminen sosiaalitieteissä. *Janus* 14 (2), 127–138. https://journal.fi/janus/article/view/50369/15219 [14.1.2021].

Forsberg, Hannele, Siiroinen, Mari & Vilkuna, Maria 2009: Third person plural in Finnish dialects: impersonal and personal uses. Workshop on Person, Impersonality and Verb Inflection in Finnic Languages and Dialects, November 12–13 2009, Helsinki. http://www.kotus.fi/files/1300/Absracts_FinDiaSyn_Workshop_2009.pdf [10.2.2020].

Gibson, Edward, Futrell, Richard, Piantadosi, Steven T., Dautriche, Isabelle, Mahowald, Kyle, Bergen, Leon & Levy, Roger 2019: How efficiency shapes human language. *Trends in Cognitive Science* 23 (5), 389–407. https://doi.org/10.1016/j.tics.2019.02.003.

Greenberg, Joseph H. 1966: Some universals of grammar with particular reference to the order of meaningful elements. Teoksessa Joseph H. Greenberg (toim.) *Universals of Language*, 2. laitos, 73–113. Cambridge: MIT Press.

Gries, Stefan Th. & Hilpert, Martin 2010: Modeling diachronic change in the third person singular: A multifactorial, verb- and author-specific exploratory approach. *English Language and Linguistics* 14 (3), 293–320. https://doi.org/10.1017/S1360674310000092.

Hawkins, John A. 2014: *Cross-linguistic Variation and Efficiency*. Oxford: Oxford University Press.

Helasvuo, Marja-Liisa & Laitinen, Lea 2006: Person in Finnish. Teoksessa Lyle Campbell & Marja-Liisa Helasvuo (toim.) *Grammar from the Human Perspective: Case, Space and Person in Finnish*, 173–207. Amsterdam: John Benjamins.

Häkkinen, Kaisa 1994: *Agricolasta nykykieleen. Suomen kirjakielen historia*. Porvoo: WSOY.

Ikola, Osmo (toim.) 1985: *Lauseopin arkiston opas*. Lauseopin arkiston julkaisuja 1. Turku: Turun yliopisto.

Itkonen, Terho 1989: *Nurmijärven murrekirja*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Itkonen, Terho 1983: Välikatsaus suomen kielen juuriin. *Virittäjä* 87 (3), 190–229. https://journal.fi/virittaja/article/view/37783 [19.1.2021].

Johnson, Paul C. D. 2014: Extension of Nakagawa & Schielzeth's R^2_{GLMM} to random slopes models. *Methods in Ecology and Evolution* 5, 944–946. https://doi.org/10.1111/2041-210X.12225.

Karlsson, Fred 1977: Syntaktisten kongruenssijärjestelmien luonteesta ja funktioista. *Virittäjä* 81 (4), 359–391. https://journal.fi/virittaja/article/view/36899 [12.1.2021].

Karlsson, Göran 1966: Eräitä tilastollisia tietoja subjektin ja predikaatin numeruskongruenssista suomen murteissa. *Sananjalka* 8, 2–23.

Laitinen, Lea 2005: Hän, the third speech act pronoun in Finnish. Teoksessa Ritva Laury (toim.) *Minimal Reference. The Use of Pronouns in Finnish and Estonian Discourse*, 75–106. Studia Fennica Linguistica 12. Helsinki: Suomalaisen Kirjallisuuden Seura. https://doi.org/10.21435/sflin.12.

Lehikoinen, Laila 2000: Uhkaako puhekieli kirjoitetun yleiskielen subjektin ja predikaatin kongruenssia. *Virittäjä* 104 (2), 246–260. https://journal.fi/virittaja/article/view/40005 [12.1.2021].

Lüdecke, Daniel 2019: sjPlot: Data Visualization for Statistics in Social Science. R package version 2.8.2. https://CRAN.R-project.org/package=sjPlot. https://doi.org/10.5281/zenodo.1308157.

Mielikäinen, Aila 1984: Monikon 3. persoonan kongruenssista puhekielessä. *Virittäjä* 88 (2), 162–175. https://journal.fi/virittaja/article/view/37845 [12.1.2021].

Nakagawa, Shinichi & Schielzeth, Holger 2013: A general and simple method for obtaining R^2 from Generalized Linear Mixed-effects Models. *Methods in Ecology and Evolution* 4, 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x.

Nakagawa, Shinichi, Johnson, Paul C. D. & Schielzeth, Holger 2017: The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface* 14, 134. https://doi.org/10.1098/rsif.2017.0213.

Nichols, Johanna 1992: *Linguistic Diversity in Space and Time*. Chicago: University of Chicago Press.

Priiki, Katri 2017: *Hän, se, tää vai toi? Vuorovaikutussosiolingvistinen tutkimus henkilöviittauksista Kaakkois-Satakunnan nykypuhekielessä*. Turun yliopiston julkaisuja 432. Turku: Turun yliopisto. https://www.utupub.fi/bitstream/handle/10024/130858/AnnalesC432Priiki.pdf?sequence=2&isAllowed=y [12.1.2021].

R Core Team 2018: *R: A Language and Environment for Statistical Computing*. Wien: R Foundation for Statistical Computing. http://www.R-project.org/ [12.1.2021].

Ros, Idoia, Santesteban, Mikel, Fukumura, Kumiko & Laka, Itziar 2015: Aiming at shorter dependencies: the role of agreement morphology. *Language, Cognition and Neuroscience* 30 (9), 1156–1174. https://doi.org/10.1080/23273798.2014.994009.

Sarkar, Deepayan 2008: *Lattice: Multivariate Data Visualization with R*. New York: Springer.

Sinnemäki, Kaius 2014: A typological perspective on Differential Object Marking. *Linguistics* 52 (2), 281–313. https://doi.org/10.1515/ling-2013-0063.

Tervola, Maija 2019: *Maahanmuuttajalääkärien suomen kielen taito ja kielitaitotarpeet: Erityisalan kielitaidon näkökulma lääkärin työhön*. Tampereen yliopiston väitöskirjat 92. Tampere: Tampereen yliopisto. http://urn.fi/URN:ISBN:978-952-03-1164-3.

Wickham, Hadley 2016: *Ggplot2: Elegant Graphics for Data analysis*. New York: Springer-Verlag.

Zipf, George 1949: *Human Behavior and the Principle of Least Effort*. Cambridge, Massachusetts: Addison–Wesley.

About the authors

Kaius Sinnemäki is Associate Professor in Quantitative and Comparative Linguistics at the University of Helsinki and the director of the ERC-funded project "Linguistic Adaptation: Typological and Sociolinguistic Perspectives to Language Variation". He works mostly on language comparison, language variation and universals, and language and religion.
E-mail: kaius.sinnemaki@helsinki.fi.

Viljami Haakana (MA) worked as a research assistant at the University of Helsinki. He started his studies of General Linguistics in 2014 and received his MA degree in March 2021. He was interviewed as "Researcher of the Month" for The Language Bank of Finland in October 2019 because of his research on this article. E-mail: viljami.haakana@gmail.com.

Språkets funktion

Juhlakirja Urpo Nikanteen 60-vuotispäivän kunniaksi

Festskrift till Urpo Nikanne på 60-årsdagen

Festschrift for Urpo Nikanne in honor of his 60th birthday

Toimittajat

Leena Maria Heikkola
Geda Paulsen
Katarzyna Wojciechowicz
Jutta Rosenberg

Åbo Akademis förlag | Åbo Akademi University Press
2021

CIP Cataloguing in Publication

Språkets funktion
Juhlakirja Urpo Nikanteen 60-vuotispäivän kunniaksi
Festskrift till Urpo Nikanne på 60-årsdagen
Festschrift for Urpo Nikanne in honor of his 60th birthday
Toimittajat: Leena Maria Heikkola,
Geda Paulsen, Katarzyna Wojciechowicz &
Jutta Rosenberg.
Åbo: Åbo Akademis förlag, 2021
ISBN 978-952-12-4061-4

Pärm: Katarzyna Wojciechowicz
Pärmbild: Jeremy Lapak on Unsplash
Ombrytning: Jutta Rosenberg

Åbo Akademis förlag
E-post: forlaget@abo.fi

ISBN 978-952-12-4061-4
ISBN 978-952-12-4062-1 (digital)

Åbo 2021